

# LGD APPENDIX 2020

LARGE CORPORATE BORROWERS

**MAY  
2020**

**APPENDIX  
DATABASE & METHODOLOGY**

## CONTENTS

|  |           |
|--|-----------|
| <b>1. DATA DEFINITIONS AND STATISTICAL METHODS</b> | <b>3</b>  |
| <b>2. REFERENCE DATA SET (RDS)</b>                 | <b>7</b>  |
| 2.1. WHY AND HOW TO CREATE AN RDS?                 | 7         |
| 2.2. REPRESENTATIVENESS OF DATA                    | 7         |
| 2.3. ELEMENTS OF THE RDS IN THIS STUDY             | 8         |
| <b>3. ABOUT GLOBAL CREDIT DATA</b>                 | <b>11</b> |
| 3.1. INTRODUCTION                                  | 11        |
| 3.2. MEMBERSHIP                                    | 13        |
| 3.3. GOVERNANCE                                    | 13        |
| <b>4. DATA STANDARDS</b>                           | <b>14</b> |
| <b>5. THE LGD/EAD DATABASE</b>                     | <b>18</b> |
| 5.1 OVERVIEW OF THE DATA                           | 18        |
| 5.2. PURPOSE OF DATABASE                           | 20        |
| 5.3. TABLES AND FIELDS                             | 20        |
| 5.4. MANDATORY VS OPTIONAL FIELDS                  | 22        |
| 5.5. INPUT VS CALCULATED FIELDS                    | 22        |
| <b>6. THE DATA QUALITY PROCESS</b>                 | <b>23</b> |
| 6.1. GCD DATA QUALITY                              | 23        |
| 6.2. STEPS IN THE DATA SUBMISSION                  | 26        |
| 6.3. DATA IMPROVEMENT                              | 29        |
| 6.4. DATA CLEANING                                 | 30        |

# 1. DATA DEFINITIONS AND STATISTICAL METHODS

## LOSS GIVEN DEFAULT (LGD)

Global Credit Data (GCD) members do not provide LGDs as an input field to the database but the underlying raw information such as outstanding amount at default or resolution and cash flows during the default. GCD calculates realised LGDs by following different methodologies (see section 5.5 for more details on input vs calculated fields). Internal methodology as well as regional regulatory requirements might result in different calculation methods which member banks are able to realise on the raw data they receive back. All variants follow the same basic definition

$$\text{LGD} = \frac{\text{Economic Loss}}{\text{Default Amount}}$$

For this report the following LGD variations are used. LGD refers to the calculation where the discounted cash flows are used. The discount rate used is the risk-free rate, more concretely the 3 months EURIBOR as at the default date. Nominal LGD is calculated in the same way but using nominal, undiscounted cashflows. All LGDs are floored at 0% and capped at 150%. The LGDs are either calculated on obligation level or aggregated on obligor level.

A detailed composition of economic loss and the default amount is given below.

## ECONOMIC LOSS

### = *Default Amount*

The calculation of Default Amount is explained below.

### + *Costs*

Global Credit Data captures direct costs including legal expenses, administrator or receiver fees, liquidation expenses and other external workout costs. These cost types are collected as separate items. Internal costs are not included.

### - *Recoveries*

The following recoveries are collected separately:

*Principal payments;*

*Interest payments;*

*Recorded book value in case the bank repossesses a collateral. It is the amount with which the credit obligation of the obligor has been diminished and which has been recorded as an asset on the balance sheet of the institution;*

*Fees and commissions received which are recoveries on extra fees and commissions charged to the obligor post default on additional services outstanding amount at resolution;*

*The outstanding amount at resolution can only be greater than 0 in case the borrower returns to a non defaulted status. In this case it can be fairly assumed that the borrower will be able to pay back its obligation and therefore the amount is treated as if it were a recovery.*

Note that post resolution payments are collected by Global Credit Data but not included in the LGD calculation method used here.

## DEFAULT AMOUNT

### = *Outstanding Amount at Default*

The amount of the principal outstanding plus past due interest as at the default date.

### + *Cash-out on Guarantee*

Any cash drawing on a contingent facility. Contingent facilities have by definition an outstanding amount of 0 at default date. The cash-out converts them into a cash obligation on which the LGD can be calculated.

### + *Financial Claim*

Financial Claims are the final adjustment of the exposure at default due by the obligor in default on a mark-to-market facility. It is the final claim, if any, of the bank against the obligor after netting all exposures and collaterals at their market value on date of liquidation.

### + *Advances*

Advances include additional funding extended post default with intention to help improve the borrower's financial condition as well as additional money drawn by the borrower as part of a revolving facility.

This calculation method is called LGD2 in the GCD terminology. Please see section 5.5 for more background.

## CURE

GCD member banks have agreed on the following definition of cure: A default having time to resolution < 1 year, no write-off and no collateral sale or guarantee call. All these items are collected separately as inputs in the data template and the cure is calculated by GCD.

## TIME TO RESOLUTION

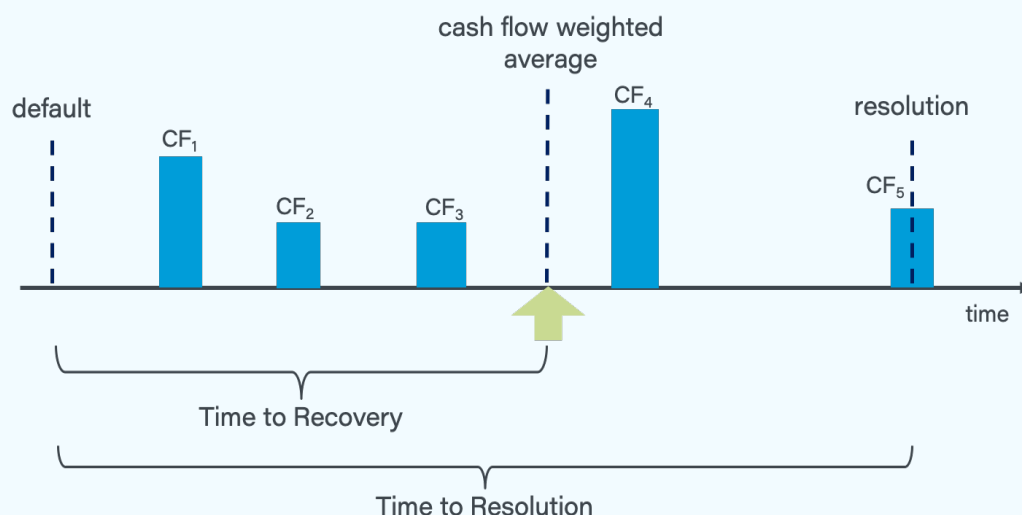
The time to resolution is the period between default and resolution. Both dates are input fields in the database.

## TIME TO RECOVERY

The term "Time to Recovery" puts a weight of the amount of the cashflow on the timing. It is defined as the cashflow weighted average period between default and cashflow. Exhibit 1 visualises the concept. The cash flow weighted time or average year of cash flow represents the weighted average of all relevant points in time between default and resolution where cash flows took place. Time to Recovery is by definition lower or equal to Time to Resolution.

## EXHIBIT 1

# CONCEPT OF TIME TO RECOVERY AND TIME TO RESOLUTION





## DEFAULT

The Basel definition of default is used. According to the GCD Data Pool Regulations all Pool participants must report their resolved defaults. Reporting unresolved defaults is recommended but optional.

## DEFAULT DATE

The date at which a borrower has been recorded as a default according to the Basel default definition

## INDUSTRY

Banks use a variety of industry codes (such as NAICS, NACE, SIC, etc). GCD member banks have agreed on a set of industry groups that banks map their internal, typically very granular, industry types to. GCD provides banks with mapping tables for the most common industry types. Therefore, consistency among banks is ensured.

## LARGE CORPORATES

Large Corporates are defined according to the Basel rules as a class of corporate exposures of € 1 million or more at group level where the reported sales for the consolidated group, of which the corporate is a part, are at least € 50 million and which is not identified in one of the five specialised lending classes, as described in paragraph 218 and paragraph 219 of the Basel II Accord.

## LOAN TO VALUE

**Loan to Value (LTV)** refers to the ratio of the outstanding amount of a loan to the value of the collateral at the default date. GCD does not collect the LTV as an input field but banks can calculate it based on the outstanding amount of the loan and the collateral value at different points in time.

## REFERENCE DATA SET

Reference Data set (RDS) refers to the data set after application of filters which is used for the analysis.

## REGION

The GCD data set offers country information on several levels (country of residence, country of jurisdiction, collateral country of jurisdiction). The items are collected on country level and returned on country level unless there are less than three banks providing data in a certain country. In this case the countries are aggregated to regions until the minimum requirement of three banks are met. This rule was established to protect anonymity of the lender identity. In this report country information is aggregated on regional level. The regions displayed are

- Africa & Middle East
- Asia & Oceania
- Europe
- North America
- Latin America

For real estate collateral GCD furthermore collects post codes on a "give to get" basis.

## RESOLUTION TO DATE

Generally, a default can resolve because of three reasons: first, the borrower pays back all the debt, second, the borrower returns to a non-defaulted status or third, the bank decides to stop the recovery efforts and writes off the outstanding debt (or sells it). The resolution date is an input field in the GCD database.

## RESOLVED/UNRESOLVED

Defaults are considered as 'unresolved' where banks are still expecting further cash flows. All other cases where the lending bank has closed the recovery file are considered 'resolved'. This is an input field in the database.

## SECURED/UNSECURED

Unsecured means the absence of collateral. A loan is unsecured if there is no collateral attached to it. A borrower is unsecured if none of its facilities has a collateral attached to it. Otherwise the loan/borrower is secured. It is important to note that no further requirements like a minimum LTV threshold are applied. As this definition creates a relatively homogeneous dataset for the unsecured data the focus in the analytics is put on this subsample.

GCD collects a variety of information on the underlying collateral. The following collateral types are reported separately:

- Cash/Reserves
- Marketable Securities
- Marketable Securities Treasury Bills
- Other Marketable Securities
- Non-Marketable Securities
- Accounts Receivable
- Inventory
- Cars
- Commercial Vehicles
- All Assets Charge or Debenture
- Specific Fixed Assets
- Ships
- Aircraft
- Other Object for Object Finance
- Real Estate
- Share capital of Subsidiary
- Intangibles
- Commodities Under Trade
- Oil and Gas Reserves and Mines
- Oil and Gas Reserves
- Other mining rights
- Telecom Networks
- Projects
- Life Policies

Additionally, there are specialised fields (e.g. for real estate collateral: the real estate type, class, location, post code, owner occupied status, ...) depending on the collateral type.

## PRIMARY/SECONDARY COLLATERAL

Primary is defined as secured by specifically identified collateral types

- Cash
- Marketable Securities -Treasury Bills
- Other Marketable Securities
- Ships
- Airplanes
- Real Estate
- and Other Objects for Object Finance
- Secondary is defined as all other collaterals.

## SENIORITY

Banks provide seniority on a slightly more granular level than displayed in this report.

- Super Senior
- Pari Passu
- Subordinated or Junior
- Equity

By definition a loan is always Pari-Passu unless the lender has made agreements with other lenders to "promote" or "demote" itself to Super Senior or Subordinated/Junior. If unknown the banks can provide an escape clause. The usage of the escape clauses is closely monitored in the audit and should be avoided if possible.

For simplicity, the first two categories have been grouped together to "Senior". Please note the grouping on obligor level: Borrowers are not always borrowing uniquely senior or subordinated. Occasionally a bank will provide facilities of differing seniority to the same borrower. The small number of bond and equity defaults as well as unknowns are also included here.

## STATISTICAL MEASURES: AVERAGE AND BOOTSTRAP CONFIDENCE INTERVALS

Average values are calculated in several sections of this report (e.g. LGD). These averages are always number weighted. For a better evaluation of the statistical estimation error for the averages, the bootstrap method is used.

The basic idea of bootstrapping is as follows. Assume a dataset with  $N$  independent realisations of the LGD. From this dataset, a new dataset of the same size  $N$  is generated by sampling with replacement. This procedure is repeated  $K$  times and the

mean value is calculated in each case. This results in a sample of possible mean values (of the size  $K$ ). From this sample, the 5% and the 95% quantiles of the distribution of the mean values are determined. These quantiles then serve as the confidence interval for the mean value of the LGD. The problem of potentially unreliable confidence intervals for heavy tails is corrected by BcA (bias corrected accelerated bootstraps) confidence intervals.<sup>1</sup>

The minimum sample  $N$  should be at least 10 records. The number of iterations should not be less than 1,000.<sup>2</sup> 10,000 iterations were performed for this report.

## 2. REFERENCE DATA SET (RDS)

### 2.1. WHY AND HOW TO CREATE AN RDS?

GCD provides members with the full “raw” data set when returning data to members after the validation and auditing process. No filtering or data cleansing is done. Banks are advised to create a reference data set (RDS) from the full data set which is a subset of observations from the full data set (borrower, loans, collateral) that should resemble the referenced portfolio.

The RDS can be used for modelling the credit risk of a portfolio, benchmarking a portfolio and validation or calibration of a model for a certain target (referenced) portfolio. This RDS should have two qualities: representativeness and data quality. The creation of a RDS is the key success factor for using pooled data.

### 2.2. REPRESENTATIVENESS OF DATA

Many regulators have set standards<sup>3</sup> or guidelines for data to be used in credit risk estimation models. Such guidelines cover

both internal and external data. In both cases the data used needs to be representative of the target portfolio during the proposed measurement time. Even internal data from a different time period, jurisdiction or sub-portfolio should be assessed for representativeness.

The total GCD defaulted borrower data set is composed of data from the banks who have chosen to be GCD members. These banks' geographical lending footprint, loan and borrower types as well as collateral practices are merged in the database. Due to the size and long time series of the database and the contributions from banks of many countries, the data set could be seen as broadly representative of an average bank, however more accurately it represents the average of GCD member banks, weighted towards the largest member banks who provide most data. Details of GCD member banks, including geographical footprint of the data, are given in section 3 below.

## NO STANDARD GCD RDS

The purpose of creating an RDS is to match as closely as possible the risk conditions of a target portfolio of a single bank. Therefore, a single standard RDS could not possibly suit all users. In this report GCD bases the analytics on a filtered data set which combines elements of representativeness and data quality.

Again, the capability of member banks to be able to narrowly define a data set aligned to the reference model significantly improves the comparability and validity of the benchmarking exercise and addresses benchmarking challenges such as differing portfolio composition, processes and policies, default definitions, weighting schemes and so forth.

## 2.3. ELEMENTS OF THE RDS IN THIS STUDY

The full database contains 13,617 defaulted large corporates on obligor level. 9,631 obligors remain in the reference data set after applying filters. The different elements and the reasons for filtering are explained below.

### UNRESOLVED CASES: CAN THE LGD OUTCOME BE CALCULATED?

Loss Given Default is most accurately calculated on closed (resolved) cases, where the outcome is anything from full repayment to complete loss, or something in between. Although GCD collects unresolved cases, the ultimate LGD cannot be calculated until the default is resolved. To avoid uncertainty by calculating proxies for unresolved cases, the RDS is restricted to resolved cases.

TABLE 1

## REFERENCE DATA SET CREATION (OBLIGOR LEVEL)

| FILTER                   | STAGE   | UNRESOLVED | YEAR OF DEFAULTS | SMALL DEFAULT AMOUNTS | INCOMPLETE PORTFOLIO | VALIDATION RULES | TOTAL         |
|--------------------------|---------|------------|------------------|-----------------------|----------------------|------------------|---------------|
| RAW DATA SET W/O FILTERS | INITIAL | 1,046      | 1,362            | 2,185                 | 219                  | 569              | 15,999        |
| UNRESOLVED               | BEFORE  | 1,046      | 1,362            | 2,185                 | 219                  | 569              | 15,999        |
| UNRESOLVED               | AFTER   | 0          | 1,141            | 2,045                 | 219                  | 411              | 14,953        |
| YEAR OF DEFAULT          | BEFORE  | 0          | 1,141            | 2,045                 | 219                  | 411              | 14,953        |
| YEAR OF DEFAULT          | AFTER   | 0          | 0                | 1,925                 | 219                  | 354              | 13,812        |
| SMALL DEFAULT AMOUNT     | BEFORE  | 0          | 0                | 1,925                 | 219                  | 354              | 13,812        |
| SMALL DEFAULT AMOUNT     | AFTER   | 0          | 0                | 0                     | 141                  | 248              | 11,887        |
| INCOMPLETE PORTFOLIO     | BEFORE  | 0          | 0                | 0                     | 141                  | 248              | 11,887        |
| INCOMPLETE PORTFOLIO     | AFTER   | 0          | 0                | 0                     | 0                    | 219              | 11,746        |
| VALIDATION RULES         | BEFORE  | 0          | 0                | 0                     | 0                    | 219              | 11,746        |
| VALIDATION RULES         | AFTER   | 0          | 0                | 0                     | 0                    | 0                | 11,527        |
| RDS                      | FINAL   | 0          | 0                | 0                     | 0                    | 0                | <b>11,527</b> |



## **YEAR OF DEFAULT: HOW TO AVOID THE RESOLUTION BIAS**

An important distinction should be identified and accordingly addressed with respect to cases with a short workout period when calculating LGD. Generally, a short workout period is related to lower LGD. In the most recent years short workout period cases are naturally overrepresented. Hence, including all the default years might lead to an unrealistically long-term average LGD. This is also important if cures are treated separately as per the GCD definition, cures are resolved within the first year from time of default whereas non-cure cases can exhibit a much longer time to resolution. This is known as resolution bias. Therefore, when creating an RDS it is advisable to address the resolution bias by restricting the defaults to those with a reasonable window time for workout processes to conclude. The decision on this filter, like all RDS filters, should be left to the discretion of users and be aligned to what is representative of their own portfolio. For the GCD dataset the average observed workout period is two years and the latest default year available in the LGD/EAD database is 2018. To address the resolution bias caused by cured cases, it is reasonable to restrict data points to defaults up to and including 2015.

A filter is applied on the lower end of the time series in addition to the filter on the upper end. Although the earliest entry in the GCD database dates back to 1983, for some banks it is difficult to deliver all the data elements required to identify cured cases for older defaults consistently with newer defaults. Such data may still be useful for driver analysis but the lower reported cure rate can tend to bias the resulting pre-2000 data such that the reported LGD is higher than it would have been in a full data set. In this report where the absolute level of the resultant LGD is important because long term averages are calculated, defaults that occurred prior to 2000 are excluded.

## **SMALL DEFAULT AMOUNT: ARE SMALL DEFAULT AMOUNTS RELEVANT?**

Default amounts in the GCD database range from zero (e.g. for uncalled contingent facilities) to several hundreds of millions of Euro. For an appropriate setup, banks are advised to compare the default amount structure to their internal portfolio. For this exercise default amounts below 100,000 EUR are excluded as they are deemed to be not representative of large corporate defaults.

## **INCOMPLETE PORTFOLIO: HOW TO DEAL WITH FORMER MEMBER BANK DATA**

When a member bank resigns from the association and/or from a Data Pool, the most recent defaulted years that they have submitted must be incomplete as they would no longer participate to submit/update their defaults. The incomplete data contains only cases with short time to resolution which might be affected by the resolution bias. Therefore, the last three years of data of former member banks are filtered out of the RDS.

## **VALIDATION RULES: HOW TO DEAL WITH OLDER DATA**

As described above, GCD applies a series of validation rules during the submission process which prevents inconsistent or incomplete data from being accepted automatically. This is the major data quality insurance that protects the database. The validation rules are updated and amended as required by our members for every submission. That said, some entries were integrated into the database before certain validation rules had been implemented. Those entries can still be part of the database if not updated by the member bank. GCD policy is to not remove any data as it may still contain useful information. However, for this exercise, data points with errors that affect the integrity of the database (e.g. the event date at default

must be the same for all facilities of a given borrower) or the correct calculation of LGD (e.g. balancing the cash flow between the transaction and the history table) were excluded. Due to the GCD rule that every bank must update their full data submission at least every three years, there are only a small number of entries removed in this filter step.

The remaining validation rules that are triggered deal with completeness. They check e.g. if for a certain collateral type, year of construction is given. Where data was submitted before the rule was

applied, the information is sometimes not provided. Since that data is not wrong it is included in the RDS. Correctness items check for example, if several connected fields are consistently filled: such as if for a syndicated loan a total syndicated amount and a currency are given. If they do not deal with crucial information used in this report the data is included in the RDS.

**TABLE 2**  
**REFERENCE DATA SET CREATION (OBLIGATION LEVEL)**

| <b>FILTER</b>                   | <b>STAGE</b> | <b>UNRESOLVED</b> | <b>YEAR OF DEFAULTS</b> | <b>SMALL DEFAULT AMOUNTS</b> | <b>INCOMPLETE PORTFOLIO</b> | <b>VALIDATION RULES</b> | <b>NUMBER OF FACILITIES</b> | <b>TOTAL</b>  |
|---------------------------------|--------------|-------------------|-------------------------|------------------------------|-----------------------------|-------------------------|-----------------------------|---------------|
| <b>RAW DATA SET W/O FILTERS</b> | INITIAL      | 2,998             | 2,416                   | 6,329                        | 1,515                       | 2,080                   | 6,427                       | 34,495        |
| <b>UNRESOLVED</b>               | BEFORE       | 2,998             | 2,416                   | 6,329                        | 1,515                       | 2,080                   | 6,427                       | 34,495        |
| <b>UNRESOLVED</b>               | AFTER        | 0                 | 1,836                   | 5,669                        | 1,515                       | 1,579                   | 5,754                       | 31,497        |
| <b>YEAR OF DEFAULT</b>          | BEFORE       | 0                 | 1,836                   | 5,669                        | 1,515                       | 1,579                   | 5,754                       | 31,497        |
| <b>YEAR OF DEFAULT</b>          | AFTER        | 0                 | 0                       | 5,413                        | 1,515                       | 1,462                   | 5,661                       | 29,661        |
| <b>SMALL DEFAULT AMOUNT</b>     | BEFORE       | 0                 | 0                       | 5,413                        | 1,515                       | 1,462                   | 5,661                       | 29,661        |
| <b>SMALL DEFAULT AMOUNT</b>     | AFTER        | 0                 | 0                       | 0                            | 618                         | 663                     | 3,767                       | 24,248        |
| <b>INCOMPLETE PORTFOLIO</b>     | BEFORE       | 0                 | 0                       | 0                            | 618                         | 663                     | 3,767                       | 24,248        |
| <b>INCOMPLETE PORTFOLIO</b>     | AFTER        | 0                 | 0                       | 0                            | 0                           | 525                     | 3,423                       | 23,630        |
| <b>VALIDATION RULES</b>         | BEFORE       | 0                 | 0                       | 0                            | 0                           | 525                     | 3,423                       | 23,630        |
| <b>VALIDATION RULES</b>         | AFTER        | 0                 | 0                       | 0                            | 0                           | 0                       | 3,300                       | 23,105        |
| <b>NUMBER OF FACILITIES</b>     | BEFORE       | 0                 | 0                       | 0                            | 0                           | 0                       | 3,300                       | 23,105        |
| <b>NUMBER OF FACILITIES</b>     | AFTER        | 0                 | 0                       | 0                            | 0                           | 0                       | 0                           | 19,805        |
| <b>RDS</b>                      | FINAL        | 0                 | 0                       | 0                            | 0                           | 0                       | 0                           | <b>19,805</b> |

## NUMBER OF FACILITIES PER BORROWER: HOW TO DEAL WITH FACILITY WEIGHTING EFFECTS

An additional filter for the obligation level is added. Borrowers with many loans which might cause overweighting are removed. The 60% of the full dataset for large corporates consists of one facility per borrower. About 98% of the dataset has less than 10 facilities attached to one borrower. There are outliers with over 200

loans attached and these lower the overall average LGD due to the bimodal left skewed shape of the distribution. These are verified loan constructions which relate to specific commercial circumstances including multi-family housing with one home per apartment or equipment finance with one loan per vehicle in a fleet. Inclusion of these would overweight the importance of a single collateral. Therefore, for more homogeneity, a filter on borrowers with 10 or more facilities is applied at obligation level.

## 3. ABOUT GLOBAL CREDIT DATA

### 3.1. INTRODUCTION

Global Credit Data is a not-for-profit organisation and was created by its member-banks to provide them with a collection of historical loss data, analysis and research resources, to contribute to a better understanding of credit risk. Global Credit Data promotes and focuses on the quality, standardisation and transparency of data, thereby improving banks' abilities to actively manage the credit risk of their portfolios.

Through its Methodology Committee and the active participation of its member-banks, Global Credit Data provides an international forum for exploring the intricacies of credit risk management and sharing of best practices. The organisation works on a "give to get" basis: rather than "shareholders' value". The active participation of its member-banks creates "membership value" for each member's immediate and long-term benefit.

### EXHIBIT 2

## OVERVIEW OF GCD SOLUTIONS



Global Credit Data's purpose is to help banks understand and model their credit risk by having access to shared data in three Data Pools: the LGD/EAD platform with historical loan loss information, the PD & Rating platform with expected PDs, migrations and defaults, and the benchmarking platform which pools predicted risk estimates, and the ICC platform with trade finance information.

Member banks remain owners of their own data. Global Credit Data receives a perpetual licence to use the data and to provide it to members. Along with sharing data, GCD aims to assist member-banks by facilitating confidential exchange of knowledge and fostering research. Global Credit Data is essentially "by banks for banks": the banks own the data and the association.

Global Credit Data has **five main elements**, as follows:

Firstly, GCD's Basel compliant data template, which captures the multiple relationships between borrowers, facilities, transactions, collateral and cash flows necessary for building and benchmarking Basel compliant models.

Secondly, GCD's securely-built, mature data collection and return portal with built in validations.

Thirdly, GCD's full system of data quality checks that range all the way from data input validation to in cycle submission audit, scoring and out of cycle submission audit.

Fourthly, GCD's "give to get" processing engine which produces a unique, anonymised, detailed data set for each bank, maximising the detail they can receive. This is run through a tested contract by a data agent, but the process, data and code is fully owned and controlled by Global Credit Data.

Finally, GCD's documentation and user assistance is a key element of the organisation. Global Credit Data has high quality documentation of both input and output data structures, to help users understand how to collect internal data to submit to GCD and to use the pooled data return. and. Helpdesk provides technical support for members during data submission and GCD executives assist with any questions on data usage, analysis and any other topics that members need help with.

## FACTS AND MISCONCEPTIONS:

### **Global Credit Data does:**

- Return the full databases to members (on a reciprocal basis)
- Issue explanatory general results based on the data
- Promote global standards in credit data collection and use
- Foster independent research on GCD data
- Facilitate method discussions between credit risk modelling practitioners

### **Global Credit Data does not:**

- Issue binding benchmark PD, LGD or EAD levels
- Claim to represent the banking industry
- Lobby regulators
- Produce and sell models

In the LGD/EAD database, Global Credit Data is not pooling banks' own estimates (although estimated LGDs are collected in the Benchmarking Platform). Instead of asking banks for their view of the LGD level according to their methods, Global Credit Data collects all the relevant facts relating to the default and the cash flows which occurred after default. In total GCD collects 120+ different data fields per defaulted obligor at different points in time. GCD then calculates the relevant LGD levels in a transparent and replicable way.

## INSIGHT TO PEER MARKET PRACTICES

In addition to the benefits of data pooling and data analytics, GCD membership allows banks unique insight into standard market practices on credit risk modelling. GCD facilitates method discussions between credit risk modelling practitioners in working groups and at their member conferences. The organisation runs various detailed method surveys including LGD modelling, stress testing techniques and CCAR. GCD also publishes analytical publications, for example Downturn LGD, Project Finance LGD rates and observed drivers, default correlations and cure rates. GCD also fosters academic studies and independent research on its data.

## INDUSTRY INITIATIVES

Global Credit Data works jointly with industry groups such as The Institute of International Finance (IIF), International Chamber of Commerce (ICC) and The Association for Financial Markets in Europe (AFME) to help its member banks discuss regulatory topics with regulators and management. At the request of its members, GCD has also been involved in direct discussions with the Risk Measurement Group (RMG) of the Basel Committee to explain how much data the industry has. Recent work has included: Revised standardised approach impact analysis (with IIF), LGD/LGL papers, commodities analytics (AMFE), Trade finance (with ICC), downturn analytics (AFME and IIF), RWA variation reasons LGD and PD (IIF), point in time vs through the cycle analysis (IIF) and RMG presentations and customised analytics on data volumes.

### 3.2. MEMBERSHIP

Global Credit Data membership is open to financial companies with compatible data, so that credit data shared will have strong

homogeneity and be generally representative for any member.

***The GCD Articles of Association allow membership if the following terms are met:***

- Licensed bank or similar credit institution
- Complies with Basel rules for credit risk measurement
- Supplies credit data to our data pools of sufficient quality, quantity and comparability to other data
- Approval by GCD's Board

A current list of GCD members is publicly issued on the GCD website under the following link: <http://www.globalcreditdata.org/members.html>

Members also receive a more detailed list of exactly which members have submitted data to each data pool, so that they have a better idea of exactly which peer banks submit data to a specific asset class.

All members of Global Credit Data remain owners of their own data throughout the process of working with GCD and are highly involved in controlling various levels of the data pooling process when working with the organisation. There is an intensive sharing of the best practice on data usage between members to upkeep global standards and GCD facilitates working groups on speciality topics to ensure there is a drive for innovation.

### 3.3. GOVERNANCE

Two main documents lay out the governance structure of Global Credit Data: the GCD Articles of Association and the GCD Data Pool Regulations.

Global Credit Data is governed by a Board, whose detailed responsibilities are defined in the GCD Articles of Association. The members of the Board are individuals appointed by the Global Credit Data General Assembly delegates representing the Members.



Specific activities around the collection, analysis and use of data are controlled by the Methodology Committee whose members are appointed by the Management Board. The detailed responsibilities of the Methodology Committee are further described in the GCD Articles of Association and GCD Data Pool Regulations.

## 4. DATA STANDARDS

GCD data is used by members for a variety of purposes including capital modelling under the Basel rules and Expected Loss Provision modelling under accounting rules (IFRS9, CECL). The detailed rules around collection of historical data (internal and external) vary greatly and therefore GCD has not officially adopted any one rule set. GCD data quality standards have been developed by practitioners from our member banks over the past decade to meet the requirements of regulatory, business and accounting purposes, (see section 6.1 below). The GCD philosophy is to return raw data to members so that they can deal with the data and make calculations in compliance with the rules affecting them.

Some examples of regulatory requirements are set out in this section.

### **BASEL II**

One of the founding principles for GCD was that the Basel II rules set a requirement for banks wanting to adopt the Advanced Internal Rating Based approach (AIRB) to collect and maintain the data necessary to build models. Banks complying with the Basel II §431 rule should be able to contribute their data to GCD's LGD/EAD data pool which requires at least the following level of detail:

*"Banks using the advanced IRB approach must also collect and store a complete history of data on the LGD and EAD estimates associated with each facility and the key data used to derive the estimate and the person/model responsible. Banks must also collect data on the estimated and realised LGDs and EADs associated with each defaulted facility."*

*Banks that reflect the credit risk mitigating effects of guarantees/credit derivatives through LGD must retain data on the LGD of the facility before and after evaluation of the effects of the guarantee/credit derivative. Information about the components of loss or recovery for each defaulted exposure must be retained, such as amounts recovered, source of recovery (e.g. collateral, liquidation proceeds and guarantees), time period required for recovery, and administrative costs."*

It should also be noted that §432 also "encourages" Foundation banks, those who do not have approval for their own LGD/EAD models, to collect the same data as above.

***The rule in Basel II §448 requires banks to use***

*"all relevant, material and available data..."*

### **and allows use of external pooled data**

*"A bank may utilise internal data and data from external sources (including pooled data)".*

### **The encouragement to base models on real historical data is given in Basel II §449, viz:**

*"Estimates must be grounded in historical experience and empirical evidence, and not based purely on subjective or judgmental considerations. [...]"*

### **BCBS239 FROM BIS**

In 2013 the Bank for International Settlements (Basel Committee for Banking Supervision) issued a document setting out "Principles for effective risk data aggregation and risk reporting" <http://www.bis.org/publ/bcbs239.pdf>. The principles are mandatory for G-SIBs from 2016 and in the future, should also be made mandatory for D-SIBs. The document is aimed at ensuring that banks can aggregate exposures to correctly measure risks across the entire bank as well as performing their internal and external risk reporting. However, many of the principles could equally well apply to GCD's data pooling with members.

### **The 11 principles which are applicable to banks cover all risk data, including market, credit and operational risk as well as Pillar 2 and risk management models as set out below:**

- **Principle 1: Governance** of risk data aggregation capabilities: The standard is for clear governance and standards approved at Board level, strong documentation, adequate resources and full awareness at Board level, including awareness of limitations.
- **Principle 2: Data architecture** and IT infrastructure: Maintenance of risk data should be a part of the business continuity planning process. There should be integrated data taxonomies across the group.
- **Principle 3: Accuracy and Integrity:** Risk data needs to be accurate and reliable, with controls as strong as those for accounting data. Risk data needs to be reconciled with sources and validated with a "balance between automated and manual systems".
- **Principle 4: Completeness:** All material risk data needs to be aggregated. Banks need to monitor their risk data for completeness.
- **Principle 5: Timeliness:** Risk reporting and aggregation needs to be produced in time to meet risk management needs of the bank.
- **Principle 6: Adaptability:** Risk data aggregation capabilities need to be flexible enough to meet ad hoc requests including scenario analysis or country specific analysis or risk.
- **Principle 7: Accuracy in reporting:** Risk management reports should accurately and precisely convey aggregated risk data and reflect risk in an exact manner. Reports should be reconciled and validated.
- **Principle 8: Comprehensiveness of reporting:** Reporting or risk should cover all material risks.
- **Principle 9: Clarity and usefulness of reporting:** The risk management reports should be clear and concise, easy to understand and yet comprehensive, with "an appropriate balance between risk data, analysis and interpretation, and qualitative explanations".

- **Principle 10: Frequency of reports:** Risk management reports need to be frequent enough to cover the speed at which the risks can change. During crises, reports need to be available at short notice.
- **Principle 11: Distribution of reports:** Reports need to be confidential but widely enough distributed to include all appropriate recipients.

In addition to the above principles for banks, there are some principles for supervisors to follow when reviewing banks in this regard. **GCD has identified that all the above principles, especially 1 to 4, are broadly applicable, by analogy, to GCD's LGD/EAD data collection, aggregation, returns and reports. Specific comments on GCD's compliance with principles 1 to 4 are covered below:**

- **Principle 1: Governance of risk data aggregation capabilities:** GCD's data pool rules, aggregation rules and return standards are well documented. Control through the Methcom and its subcommittees and ultimately by members in the General Meeting, is clear.
- **Principle 2: Data architecture and IT infrastructure:** GCD uses well-documented and defined data architecture. There is an integrated data taxonomy across all 3 data platforms.
- **Principle 3: Accuracy and Integrity:** GCD's data integration is performed in an accurate and reproducible way. The process accurately transforms the new input data into the total data pool, without losing any data points. GCD's principles of not amending, adding or removing any information, but only allowing such changes to be made by the members, makes the entire process reconcilable. As an example, 900 defaulted borrowers in

Facility Asset Class 4 may exist. When banks add new data say with 25 amended borrowers and 50 new borrowers and ask to move 10 borrowers to another Facility Asset Class, then the total number must reconcile to 940 resulted borrowers. This delta analysis is carried out internally. Banks know exactly which data they have put in and changed and are able to run this calculation for themselves to counter check GCD.

Accuracy of calculation is ensured by the documentation and transparency of the calculations made. In the LGD/EAD database every calculation made by GCD, e.g. LGD or Time to Resolution, can be replicated by the banks who receive the data in return. GCD is not able to confirm the accuracy of the data input by the members, as only their auditors can do this. However, GCD does perform both hard technical rule checks and softer expert reasonableness checks on the data input and then rejects the non-compliant data and asks banks to re-input their data, (see later sections).

- **Principle 4: Completeness:** An interpretation of what completeness means for GCD's LGD/EAD data could be that GCD should receive a complete delivery of all defaulted cases in each facility asset class from each of the banks delivering to that facility asset class. In addition, the data fields for each of the contributed borrowers and loans should be complete. GCD ensures this completion by setting certain fields as mandatory and not accepting data which does not contain sufficient information. This is discussed in this paper in section 6.1.

## ECB'S TRIM GUIDELINES

The ECB's targeted review of internal models, or TRIM, is a project to assess whether the internal models currently used by banks comply with regulatory requirements, and whether they are reliable and comparable. [https://www.bankingsupervision.europa.eu/ecb/pub/pdf/trim\\_guide.en.pdf](https://www.bankingsupervision.europa.eu/ecb/pub/pdf/trim_guide.en.pdf)

**TRIM guidelines cover the following data quality dimensions:**

- a. Completeness - Values are present in the attributes that require them
- b. Accuracy - Data is substantially error free
- c. Consistency - A given set of data can be matched across different data sources of the institution
- d. Timeliness - Data values are up to date
- e. Uniqueness - Data is free from any duplication from filters or other transformations of source data
- f. Validity - Data is founded on an adequate and rigorous classification system
- g. Availability/Accessibility - Data is made available to the relevant parties
- h. Traceability - The history, processing and location of the data under consideration can be easily traced

These dimensions are expected to apply to internal, external and pooled data. (Section 91.2 104 (a).

**TRIM specifically calls for the use of external data in benchmarking in paragraph 57 (e) 1:**

*"Additional tests that should be performed on a periodic basis are:*

*(i) Benchmarking analyses:*

- *the bank should carry out comparisons with representative, comparable, external up-to-date data sources, and in particular with low-default-portfolios (Article 185(c) of the CRR)."*

## US FED BENCHMARKING GUIDELINES

The US Federal Reserve in its supervisory regulation letters (SR11-7) makes clear demands on the use of external data (and models) for benchmarking in <https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf>

*"Benchmarking is the comparison of a given model's inputs and outputs to estimates from alternative internal or external data or models. It can be incorporated in model development as well as in ongoing monitoring. For credit risk models, examples of benchmarks include models from vendor firms or industry consortia and data from retail credit bureaus. Pricing models for securities and derivatives often can be compared with alternative models that are more accurate or comprehensive but also too time consuming to run on a daily basis. Whatever the source, benchmark models should be rigorous and benchmark data should be accurate and complete to ensure a reasonable comparison."*

## EBA GUIDELINES ON PD ESTIMATION, LGD ESTIMATION AND TREATMENT DEFAULTED EXPOSURES

The EBA guidelines deal with the data requirements in section 4.2. (They are available under the following link:

<https://www.eba.europa.eu/documents/10180/2033363/Guidelines+on+PD+and+LGD+estimation+%28EBA-GL-2017-16%29.pdf/6b062012-45d6-4655-af04-801d26493ed0>

## 5. THE LGD/EAD DATABASE

### 5.1. OVERVIEW OF THE DATA

GCD has built up the world's largest non-retail bank loan loss database with over 195,000 defaulted facility observations totalling over €550 billion in all non-retail Basel asset classes. The Global Credit Data LGD/EAD platform, the largest of GCD's databases, was created in 2005 and has grown rapidly ever since. It now gathers detailed information on over 100,000 of defaulted counterparties, particularly data on cash flows and collateral. The driving principles and objectives of the Global Credit Data databases are:

- **Confidentiality**  
GCD ensures data is fully anonymised
- **Comparability**  
The data GCD utilises is only from banks with common definitions
- **Data Quality**  
GCD ensures the highest standard of data quality through a rigorous process of validations, auditing and scoring
- **Granularity**  
A full database is always returned to member banks, not just the aggregate
- **Reciprocity**  
Member banks must contribute to the databases before they can benefit from them by asset class and year
- **Best practice sharing**  
GCD utilises method workshops, surveys and HPE
- **Research standards**  
There is always a common basis for shared analysis and research

The data collected by GCD comes from over 50 banks across the span of 20 years. Most defaults occurred from the year 2000 forwards. Below is an infographic of the loans in the LGD/EAD database including asset class alongside the number of

defaulted loans and as expected, SME data is the most numerous.

Global Credit Data returns the complete database of detailed transactions to member banks, excluding borrower names and Lender ID's. Members also receive some pre-calculated metrics (EAD, LGD, etc.) but are encouraged to calculate their own, using their own methods. Members are then free to sample the database, clean the data, calculate their own metrics and extract drivers.

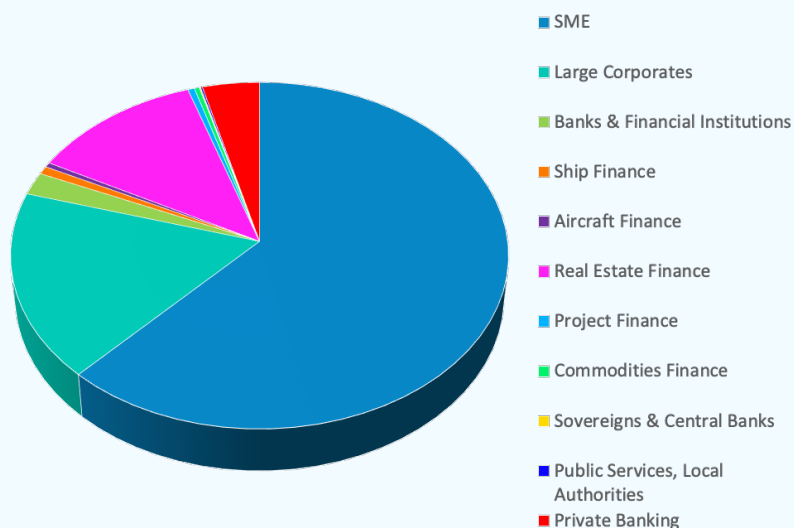
Each member chooses the data pools (asset classes) to which it contributes. The most popular is the large corporate asset class, comprised of loans to corporates where the borrower group has a turnover of €50m or less. 58 banks in total have contributed to this asset class, largely driven by the fact that nearly all members have a large majority of their lending books by volume taken up by large corporate lending. Next most popular are Banks and Financial Companies, SME and Real Estate Finance, again driven by the ubiquity of these assets in loan books. More specialised financing such as Aircraft or Shipping Finance comprise a more limited club of around 20 or less lenders over time.



### EXHIBIT 3

## FACILITIES IN THE LGD/EAD DATABASE

| FACILITY ASSET CLASS               | NUMBER OF FACILITIES |
|------------------------------------|----------------------|
| SMALL/MEDIUM ENTERPRISE (SME)      | 122,035              |
| LARGE CORPORATES                   | 34,495               |
| BANK & FINANCIAL COMPANIES         | 4,140                |
| SHIP FINANCE                       | 1,464                |
| AIRCRAFT FINANCE                   | 849                  |
| REAL ESTATE FINANCE                | 23,839               |
| PROJECT FINANCE                    | 912                  |
| COMMODITIES FINANCE                | 691                  |
| SOVEREIGNS & CENTRAL BANKS         | 173                  |
| PUBLIC SERVICES, LOCAL AUTHORITIES | 254                  |
| PRIVATE BANKING                    | 7,887                |
| <b>TOTAL</b>                       | <b>196, 739</b>      |

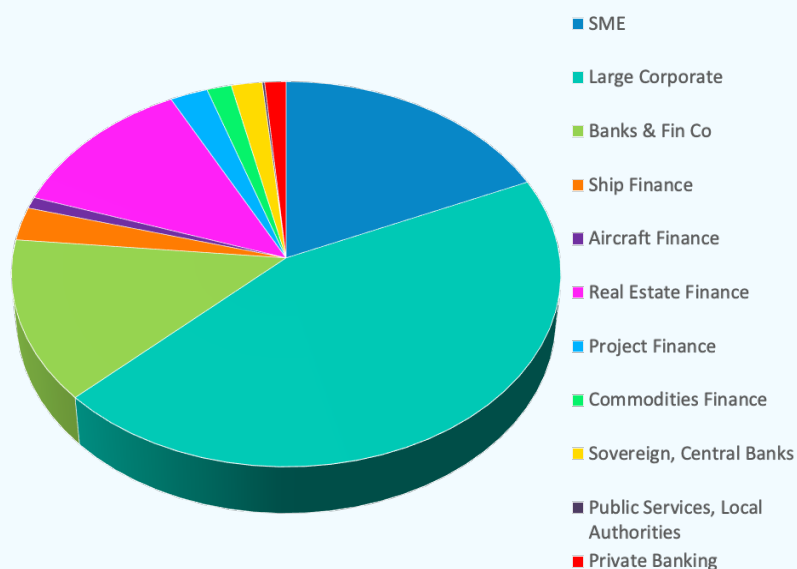


By number of loans, the SME asset class receives most data understandably, with Large Corporates and Real Estate Finance in second and third place. In terms of exposure in the LGD/EAD database, the figure below highlights exposure of defaulted loans in the LGD/EAD database in conjunction with borrower asset class. As expected the largest defaulted loan volumes come from Large Corporations.

### EXHIBIT 4

## EXPOSURE IN THE LGD/EAD DATABASE

| FACILITY ASSET CLASS               | EXPOSURE (IN MN EUR) |
|------------------------------------|----------------------|
| SMALL/MEDIUM ENTERPRISE (SME)      | 103,106              |
| LARGE CORPORATE                    | 251,986              |
| BANKS & FINANCIAL COMPANIES        | 76,729               |
| SHIP FINANCE                       | 15,686               |
| AIRCRAFT FINANCE                   | 5,365                |
| REAL ESTATE FINANCE                | 69,292               |
| PROJECT FINANCE                    | 13,623               |
| COMMODITIES FINANCE                | 8,846                |
| SOVEREIGN, CENTRAL BANKS           | 11,125               |
| PUBLIC SERVICES, LOCAL AUTHORITIES | 857                  |
| PRIVATE BANKING                    | 7,684                |
| <b>TOTAL</b>                       | <b>564,301</b>       |



## 5.2. PURPOSE OF DATABASE

The purpose of the database is to develop banks' understanding of both the quality of their data in comparison to other banks and the risk involved in the loans. Global Credit Data member banks receive back a database of detailed transactions, excluding borrower names and Lender ID's. Due to the "give to get" rules applying, members only receive back data for the years of default and asset classes which they submitted. Members are then free to sample the database, clean the data, calculate their own metrics and extract drivers. When the data set is returned to member banks, each member can then slice and dice it to produce matching portfolios before finally creating a representative reference data set which is the key success factor when using pooled data.

## 5.3. TABLES AND FIELDS

The structure of the database reflects the full complexity of the legal relationship between a bank lender and a borrower. It is designed to deal with the simplest through to the most complex deals. The types of complexity covered include:

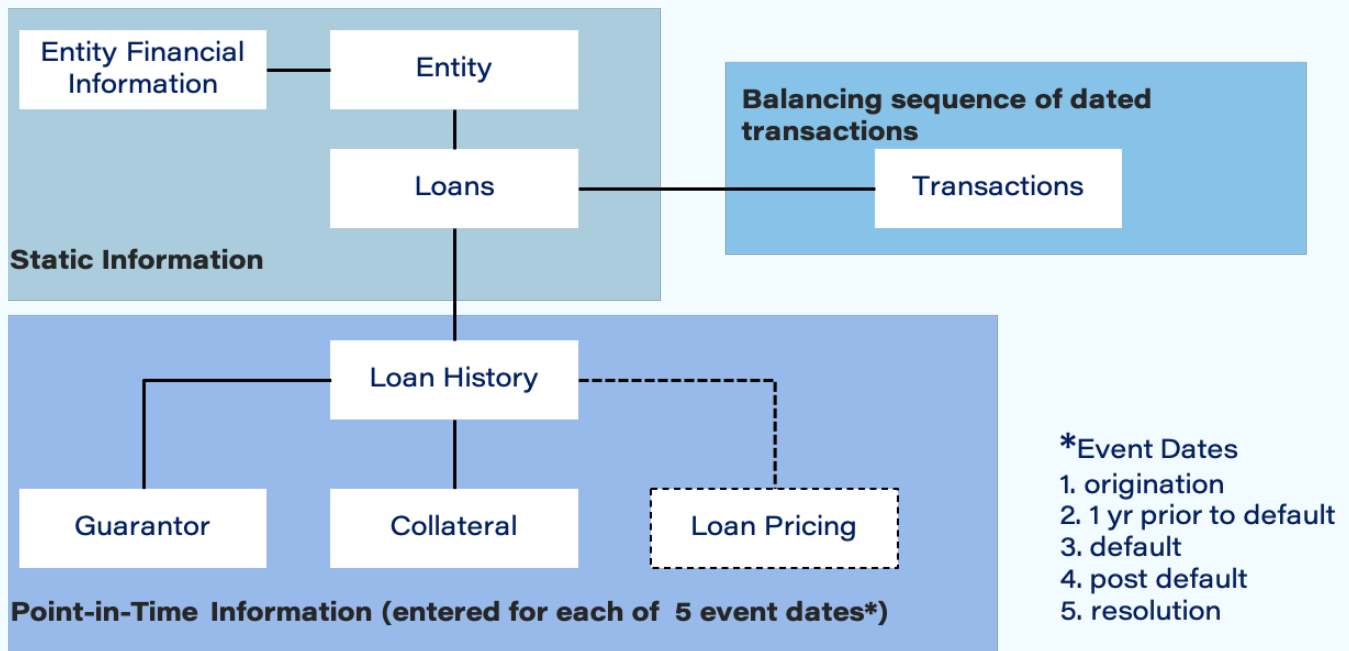
- Single company borrower with multiple facilities (loans, commitments and off-balance sheet)
- Each loan or group of loans having single or multiple full or partial guarantors
- Each loan or group of loans being fully or partially secured by multiple collaterals

In addition, the information around loans, guarantors, collateral and pricing is time stamped so that the changing loan limits, collateral values and guarantor coverage can be reflected at different points in the life of the loan from origination through to resolution. The data that is used in the LGD/EAD database is recognised as the industry standard and is used by many banks for their own internal data collection to build a predictive model of lending for the future.

The data model consists of eight interlocking tables. The entity, the loan and the entity financial table contain the static information e.g. country of residence in the entity table or facility type in the loan table. The following tables capture information that changes over time. The loan history contains among others the information on limits and outstanding amounts. Guarantor and Collateral information is collected in separate tables. The loan pricing contains information on the interest rates. Finally, in the transaction table the cash flows that occurred between default and resolution (or post default date for unresolved) are collected each separated by date, type and source of payment.

## EXHIBIT 5

# THE LGD/EAD DATA MODEL – 8 INTERLOCKING TABLES

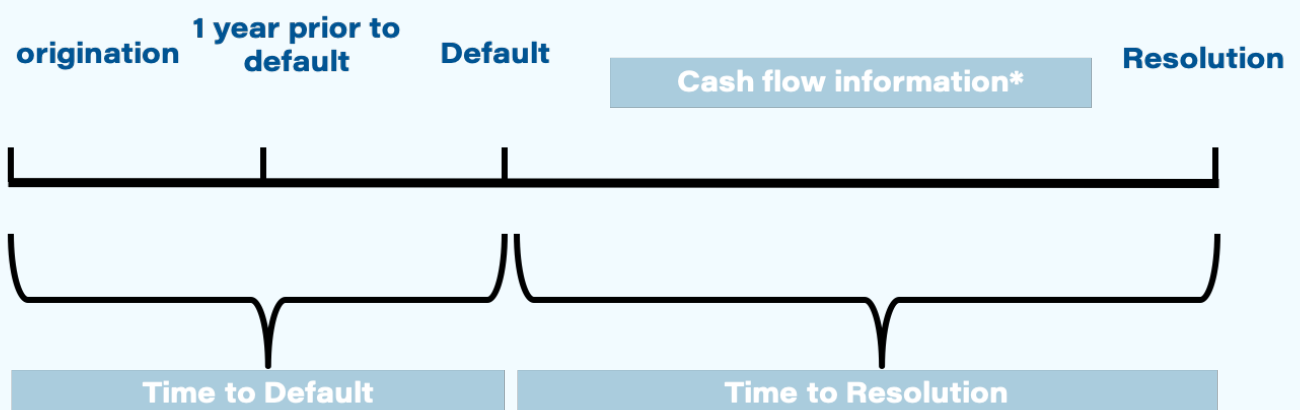


Data is collected from the point of origination, 1 year prior to the default, the point of the default and later time stamped data (unresolved cases only) up to and including the point of resolution (resolved cases only). Cash flow and accrual transactions are collected in relation to loans, collateral

and guarantors and are time stamped to exact dates, sources and purposes. This is key for discounting of cash flow to produce economic LGDs and for calculating the progress of defaulted loan cases over time.

## EXHIBIT 6

# THE LGD/EAD DATABASE STRUCTURE



## 5.4. MANDATORY VS OPTIONAL FIELDS

Each of Global Credit Data's databases has tables and fields for members to fill with data. Most fields are mandatory. Some fields have extra information which members may find difficult or legally challenging to provide, such as pricing information or borrower financial statistics. These fields are therefore optional and provide extra information for the database that is not necessarily vital for it to function correctly.

Depending on the use of the database, the optional fields may or may not be relevant. As an example, data on limits and exposures at origination of the loan is optional, while the same information at date of default is mandatory. A bank wanting to use the data for historical analysis or to bring in a time to default driver would wish to use the origination data, however a bank

seeking to use it for LGD level calibration may disregard this altogether.

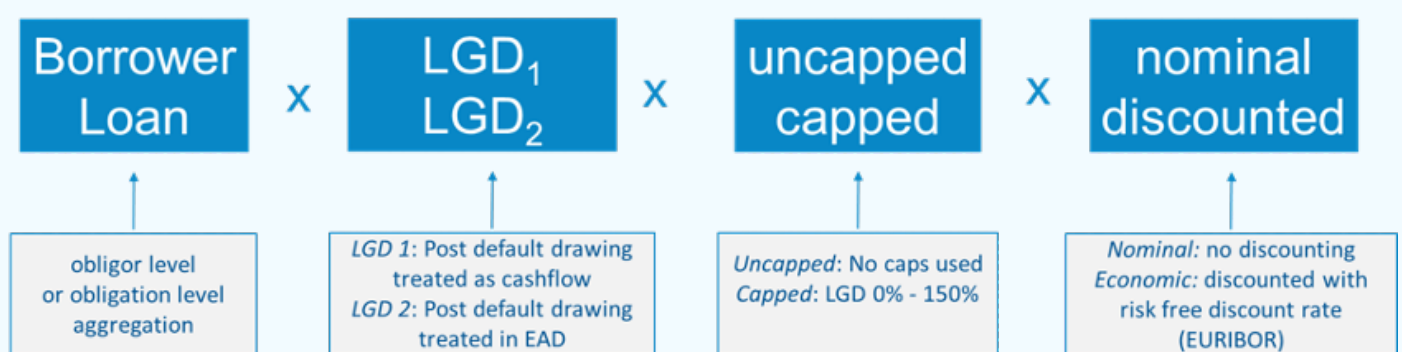
## 5.5. INPUT VS CALCULATED FIELDS

The GCD data collection covers basic information as inputs. Based on these inputs GCD calculates certain fields as outputs. These fields include:

- Replacing country codes with regions, if they do not meet critical mass rules to ensure anonymity
- Replacing identifiers with a global uniform set to ensure anonymity
- Calculation of variables such as LGD, Recovery Rate, Cures based on a methodology agreed by the GCD members. Collecting all the relevant facts that relate to the default and the cash flows which happened after default, enables GCD and users of the data to calculate their own view of EAD and LGD, according to differing methodologies. For easy use of the GCD data the members have agreed on a certain methodology.

## EXHIBIT 7

# DIFFERENT LGD CALCULATIONS IN THE GCD DATABASE



GCD provides the following options for LGD and Recovery Rates.

Firstly, there are two different aggregation levels. All facilities can be treated separately or aggregated at borrower level. GCD recognises that there are different aggregation levels used by its members and therefore provides calculations on both levels.

Secondly, GCD provides an option on how to treat advances after default. Based on whether members include or do not include advances in their EAD/CCF estimations, they can consistently use LGD1 or LGD2. In LGD1 the advances are included in the loss calculation (numerator) only. In LGD2 advances are included in the default amount calculation (denominator) as well.

Furthermore, the LGD range is considered. GCD provides an uncapped, unfloored option but also a version where a floor of 0% and a cap of 150% is applied.

***Finally, there are the following options for nominal or discounted LGDs:***

- Nominal - An LGD calculation without discounting
- Risk-Free - Discounted with 3 months EURIBOR
- Other discount rate - Use raw cash flows to calculate own LGDs.

Contractual interest rates are collected as well and banks wishing to use them for discounting do have the option. Banks can either provide the base rate plus spread or the total spread charged to the loan before it was transferred to default.

As different regulators have established different discount rate requirements, the LGDs for the defaults in the GCD database can also be calculated based on the cash flows and individual discount rates. A higher discount rate has a more pronounced effect on LGD for longer workout cases with high recovery rates, where the quantum being discounted is higher (see [GCD Discount Rate Study](#) for more details).

## 6. THE DATA QUALITY PROCESS

### 6.1. GCD DATA QUALITY STANDARDS

GCD adheres to several defining data quality principles built in from when it started in 2004. These principles are used on all pooled data to ensure the data is of the highest quality. GCD carefully screens banks data on entry through validation, audit and scoring. On top of this, banks joining GCD have greatly improved their data quality by cleaning and improving their existing internal data to reach GCD data quality standards. This includes a requirement to fully re-submit all data every three years which helps banks keep up to date with improving validations, new fields and changed definitions.

#### DEFINITION OF DATA QUALITY

An initial view of data quality might be that it concerns accuracy, i.e. whether the data accurately reflects the outcome of the default case where the times and dates of cash flows and balances are correctly reported. From the viewpoint of the user

of the data, this is necessary but not nearly sufficient. The user requires that the data is:

- Accurate (every element is correctly recorded)
- Complete (where each case presents the full story)
- Information rich (covering main drivers of risk)
- Representative (capable of being cut to a representative set)
- Unbiased (with respect to the area of use).

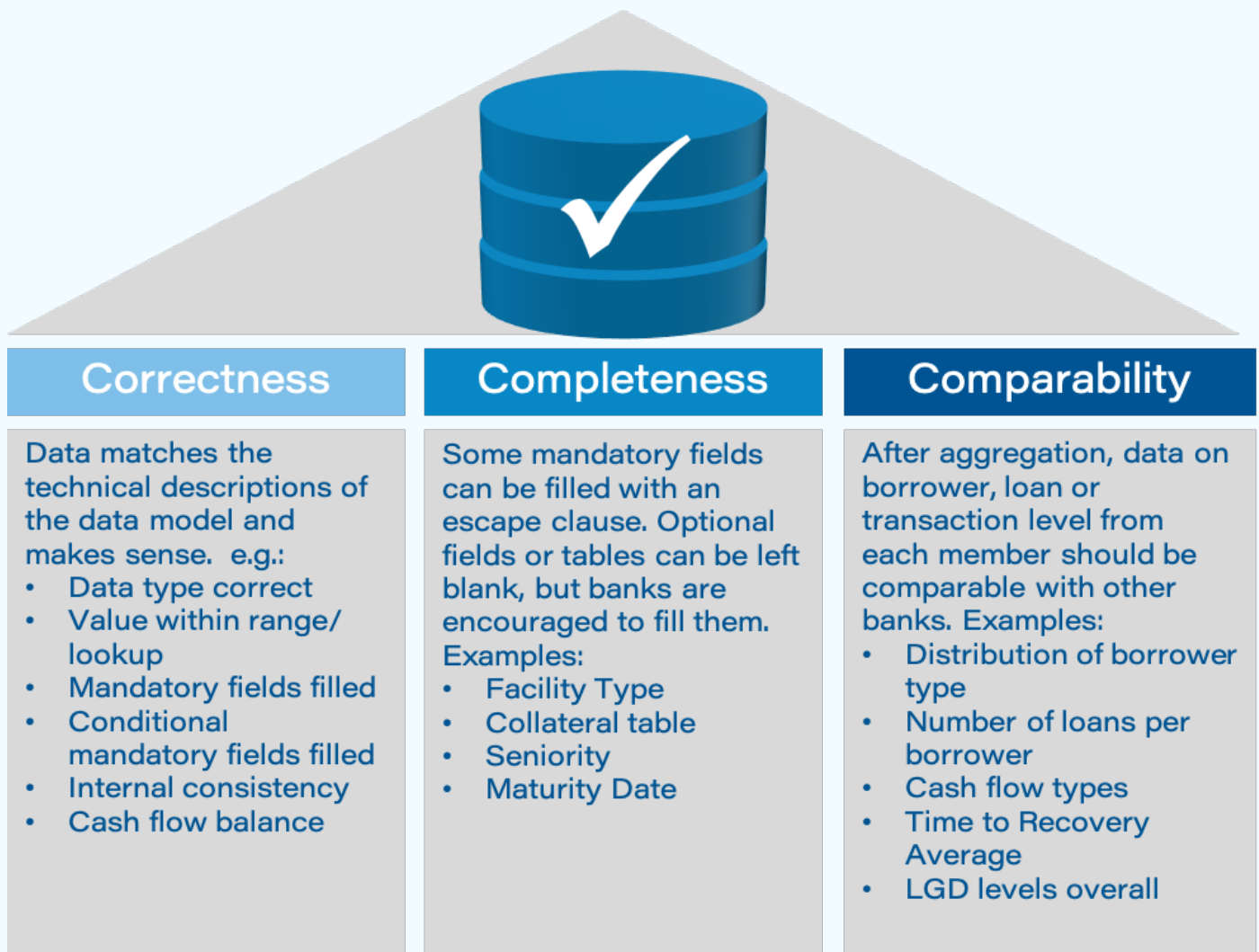
Whether the data is of sufficient quality depends very much on the purpose it is being used for. For example, a user attempting to find average historical Credit Conversion Factors for EAD modelling will focus on the areas of facility type, limits and drawings prior to and at the default date whereas a user searching for average unsecured bank LGDs will focus on borrower type, post default cash flows and write-offs and presence of collateral. In each case the user will require that the above 5 conditions are satisfied for the subject data.



For simplicity, Global Credit Data has defined three pillars of data quality, which are used in the validation, audit and scoring of the data input:

## EXHIBIT 8

### THREE PILLARS OF DATA QUALITY



#### DATA QUALITY CONTROL PROCESS

*The elements for controlling data quality comprise both GCD's formal processes and coherence checks on the aggregated data outcome.*

**Data Delivery:** The first step in the GCD quality process is in the data delivery in

which the banks are only able to submit the data if it passes many strict quality tests (validation rules). The tests of course check for correct field formats but in addition to this they also perform completeness checks based on borrower, loan, collateral and guarantor level. There are also logic checks on each submission to make sure that the cash flows and accounting entries reported

add up. It is these last tests that are the most difficult for the submitting banks and it is normal for banks to have to seek completing information from the collections unit within the bank before being able to submit.

**Data Audit:** Next step in the GCD quality process is the audit, this occurs after the delivery and before return of the data. In this phase a GCD executive looks at the delivery for each bank and reviews it for reasonableness and completeness. The executive then contacts banks where necessary to request clarifications or amendments. They also issue an audit letter for each bank highlighting their weaknesses. It is only by using this experienced human expertise that data gaps can be determined and fixed. Examples include banks delivering data in which every field and every borrower passes the filters but which, when the patterns are examined, shows that they are entirely missing any reference to guarantors, for example. The bank is then asked whether in their lending policy and country they do not use guarantors. A satisfactory answer must be received.

**Data Scoring:** Then GCD moves on to a data scoring system focussing on "bad points" comprising missing "non-fatal" fields, e.g. industry code, and other common data quality issues. The number of bad points (by reference to proportions of the delivery so as not to penalise large banks) are totalled and a ranking of all banks is made. Banks are given a detailed scoring of their bad points as well as a list of scores by other banks (anonymised of course) both overall and for individual issues. In this way, a member bank can gauge the quality of the data by reference to their own known data quality and can judge the quality of individual elements with which they are concerned in their modelling. GCD uses this scoring and ranking to encourage banks to improve and to gauge the overall improvement in data quality over time.

**Cleaning during RDS:** Member banks are also advised and encouraged to examine the detailed raw data they receive in the twice-yearly data return and to produce their own "representative data set" or RDS which comprises a filtering of the data to ensure that it matches the portfolio of the bank which will be using it. This is normally done by limiting the data to only countries of interest, asset classes of interest and often particular drivers such as industry, seniority or security. During this process the final data cleaning is done such that data which is possibly wrongly classified, or which may have data errors or which may be incomplete for this analysis, is removed.

**Regular review:** Furthermore, the GCD Articles of Association make clear that the organisation has the right to remove any data from the database if the GCD executives feel that it is not delivered in good faith and indeed as an ultimate sanction, GCD has the right to expel any member. This latter sanction has never been employed however GCD has in its history had the case of a member bank voluntarily removing data of insufficient quality after discussions with the executive.

**Member data quality overview:** All GCD member banks are subject to Basel II regulatory scrutiny. Any bank which uses GCD data for either model building or benchmarking is required to prove to its regulator that the data is of good quality and is appropriately representative. GCD data and the RDS produced therefrom, has been submitted by many banks to many regulators with very good results. The regulators are aware of the default levels of their regulated banks and can see whether the bank has submitted sufficient data to GCD. GCD data's granularity, size and source makes it the best quality pooled data available for PD, EAD and LGD.

## RESPONSIBILITY FOR DATA QUALITY

Global Credit Data takes its responsibilities in line with GCD Data Pool Regulations and data quality seriously and requires its

members to do the same. Article 4A of the GCD Articles of Association sets out an assurance of data quality within the activities of the organisation. Members are obliged to ensure that data they deliver to the Association is of good quality. To support members with this, executives conduct audits to check the data before it is put in the database. Review during the audits consists of both technical and expert review, including special analytical audits. Executives have the right and the obligation to perform in cycle and out of cycle audits and share the results with the members. Based on this they decide whether data is of sufficient quality to be added to the database or not. In case of doubt the executives refer the matter to the Methodology Committee.

Each member shall carefully consider the outcome of audits and perform updates where required to deliver a high level of data quality. If the contribution of a member is not of sufficient quality executives will report this to the Methodology Committee in such a way as to ensure anonymity of the Pool-participant and objectivity of the Methodology Committee. The Data Template for each Data Pool may be amended from time to time to improve data quality and completeness. Such amendments may include new or removed fields or tables, new or removed field values and new or amended validation rules and will be approved by the Methodology Committee. As a result, it is desirable that previously submitted data is updated and resubmitted. Where possible each member of a Data Pool is required to re-submit all credit data previously submitted at least every three years. If a Pool

participant is unable to provide all or any credit data previously submitted (for example due to resource issues) it may ask the Methodology Committee for an exemption of its requirement to provide credit data for a given historical period, in which case access for such Pool participant to the credit data in the corresponding Data Pool will be restricted. If resubmitted data does not pass certain current validation rules which have changed since the time of the original submission of that data, a Pool participant can ask to bypass certain new rules in order to gain access to the data in the database. This will only count for previously submitted data, not for new data. Such Pool participant shall explain to the executives the reasons for not being able to update data to the new rules and the executives shall decide on whether to accept the data in this form or not based on the age of the original data, the overall completeness of the data and the risks of incorrectness, incoherence or bias. A Pool participant receiving a rejection to such request may seek a final decision from the Methodology Committee or a Pool participant can ask the executives to take the request to the Methodology Committee to ensure anonymity.

## 6.2. STEPS IN THE DATA SUBMISSION

GCD has a well-defined data process with strict automatic validation rules and a tough expert based auditing system. The banks submit their data either between April and May or October and November in the two submission windows. Data is submitted during these periods to allow GCD to have a complete fresh data set to return each June and December.

## EXHIBIT 9

# DATA SUBMISSION PROCESS FLOW



The data submission process begins with member banks receiving individual pre-submission packages on their data submitted so far. These customised analytics identify improvement fields for each bank. During submission, the data is checked by the validation rules in an automated, iterative process. After all validation rules have been successfully passed, the data is confirmed by the member and then separately checked by executives according to auditing standards consistently applied. Both the validation and auditing rules are set and regularly reviewed by the Methodology Committee. After the data is audited, it is then aggregated, anonymised and readied for return. At this point the data submitted by each bank, together with their previously submitted data is scored using pre-set data quality scoring.

## SUBMISSION TYPES

Banks can choose between full submission and partial updates. By definition, a first submission of a new member is a full submission and the GCD Data Pool Regulations require at least one full submission every three years. Resubmission of any existing data is done through the data portal as usual, where data is tested with exactly the same tests as for new data. The resubmitted data is checked by the audit process as usual and if accepted is then used to replace the old data as a step in the aggregation function.

### **Full Submission requires:**

- Submission of all resolved defaults present in the GCD dataset
- Submission of all unresolved defaults present in the GCD dataset
- Submission of new defaults (resolved/unresolved) if available
- Partial Updates require:
- Submission of none or some existing defaults present in the GCD dataset
- Submission of new defaults (resolved/unresolved) if available

## VALIDATION RULES

The first hurdle banks have to pass is the process in which data goes through extensive tests to ensure that data meets GCD's current data quality rules. The input checks are setup to correct any misinformation or faulty data that is entered into the web portal. There are two types of validation rules:

- **Error:** Data is incorrect, validation rules will prevent members from submitting this data
- **Warning:** Data has an unsatisfactory or unlikely value, banks are encouraged to check and correct if necessary although submission is still possible.

Member banks must first prepare their data in the 8 interlocking tables and create a submission file ready to submit using CSV, Excel or XML data formats. Banks can then log on to GCD's secure portal and upload files in a live session of validation and submission. These trial submission sessions can run over several days as banks gradually remove all the errors. Technical and methodology support is available during this period. When all errors are removed, banks can submit their data as a firm submission. This real-time validation process ensures consistency across data pool, promotes transparency and shortens the learning curve.

Global Credit Data's validation rules are contained in GCD's 'Loan Loss Database Data Input Structure' document.

## AUDIT OF THE DATA

GCD executives will then audit each bank's submission and revert with expert comments on data quality and suggestions to improve, and resubmit if necessary. Member banks then revise their data. Data is only accepted once the audit process is completed successfully.



**How do data audits happen?** A Global Credit Data executive reviews each delivery of data for reasonableness and completeness, followed where necessary, by a request for clarifications or amendments.

**How is the audit issued?** Global Credit Data issues an audit letter for each bank which highlights weaknesses and forms a data quality track record for their management.

**How are patterns of bad data determined?** Human expertise is needed to determine persistent patterns of insufficient data quality which may pass automatic filters (defaults missing any reference to guarantors, for example). Only executive experience can sense the non-reporting of some defaults (missing types of exposures, unresolved through long pending defaults, non work-out defaults, etc.). Such pattern can also be detected during specific analytics performed during the previous period resulting in additional questions and correction requests to member banks.

## IN-CYCLE DATA AUDITS

Data audits are produced to assist banks in ensuring that data is reported uniformly and follows business rules. There is a focus on data completeness and encouraging more granularity, for example with fields reported as "unknown" or with missing information such as missing collateral information.

## PRE-SUBMISSION PACKAGE INCLUDING OUT OF CYCLE AUDITS

Prior to this data submission, GCD executives prepare an individual analysis for each bank using the existing data that helps them prepare their submission data for the next cycle. The so-called pre-submission packages contain a large variety of analytics on correctness,

completeness and comparability to other banks and are the starting point for improving and completing data.

A pre-submission package is prepared before each official semi-annual submission including an audit of the full data input from each member bank using internal queries and validation rules results. 500+ items are checked including:

Audits are performed by an experienced Global Credit Data Executive with senior credit experience and a long history with the data model. Significant evidence can then be gained on data quality strengths and weaknesses of the database.

## PROCESS THE DATA

***The next step is to process the input data. GCD performs the following steps in aggregating the data before return:***

- Updating old information about default cases previously submitted (replacement functionality)
- Adding new cases
- Replacing borrower ID, Loan ID, Guarantor ID and Collateral ID with a uniform set Data Agent ID; "DA" (Anonymity)
- Adding the calculated fields
- Replacing country codes with regions, if they do not meet critical mass rules (Anonymity)

Producing details of each member's data submitted for review of the data entitlement. Banks must resubmit their existing defaults during this process by resubmitting the default with the same Entity ID as present in the GCD Dataset. During the aggregation phase, GCD's process will identify the borrower's resubmitted data based on their Entity ID and replace all existing information.

The final step of the data process is the return of the data to member banks. GCD executives review the "give to get" and authorise a list of asset classes and years to be returned to each bank called the lender entitlement. GCD creates individual return



data files for each bank and places them on the portal. After this is completed return data is available for members to download.

## SCORING

Based on its data submission each bank is assigned a single score. The scoring ranges from 0 (good) to 6 (bad). The scoring allows monitoring improvement on data quality by item, asset class and bank. 40% of the score, represents the correctness of the data; whether the data complies with the most recent validation rules. If an entry is new or resubmitted, the score is automatically 0. Another 40% of the score covers completeness checks. It scores if a bank submits optional fields or optional tables. The final 20% scores comparability. It compares banks' LGDs, Time to Recovery and Cure Rates and detects if a bank's submission is significantly different from the pool. Naturally, this must be a softer criterium as not all the banks are expected to be the same but it triggers a discussion on completeness of the submission if levels are very low or high compared to peers. The scoring is a tool where the absolute numbers are not so important but rather tracking over time shows improvement in the overall dataset as well as by individual members.

## THE MOST IMPORTANT VALIDATION RULES

To enter the GCD database data must go through 450+ validations. This real-time validation process ensures consistency across the data pool, promotes transparency, shortens the learning curve and most importantly keeps the level of data quality high. Validation rules and methodologies have been developed and constantly reviewed by the Methodology Committee. These validations are documented extensively for easy consultation during the data submission process and reconciliation with the data definitions and objectives.

Validation rules are built into every field of every table in the web portal where information goes in and gives banks automated feedback on every rule broken and all warnings about data they have either filled in or forgotten to fill in. For example, simple but crucial checks are if the format is correct (e.g. in a date field) or the correct drop-down items from the lookup tables are used. But it is also checked if related fields are consistently filled. One example is if banks have submitted that they have real estate collateral they must fill in what type of real estate it is. Some validation rules are only triggered if another field is filled with certain information. For example: The transaction table records all cash flows between default and resolution. This can be incoming cash flows (e.g. principal payments or interest payments), outgoing cash flows (e.g. principal advances) or write-offs. For entries which are incoming cash flow types another field indicating the source of the payment must be filled.

Two of the most important validation rules concern the balancing cash flow sequence. Banks deliver the outstanding loan amounts at different points in time, more concretely for the date of default and the date of resolution. For the period in between, the recovery period, the entire record of cash flows, including costs, interest, etc are reported in a different table. GCD checks if the cash flows balance to the amount of their booked loss. Only 5% underpayment and 10% overpayment is allowed. If the cash flow does not balance for a given loan, the validation rules will trigger an error and banks need to investigate and balance the transaction entries to pass the validation.

After complying with the automated validation rules that triggered an error message, banks can commit their data. Only if all errors are removed GCD accepts the data on the portal.

## 6.3. DATA IMPROVEMENT

While the validation rules act to keep out data which does not have sufficient detail or which does not balance arithmetically,

their primary aim is to highlight this to the submitting bank and give them the opportunity to improve the data. The data submission process occurs over a period of weeks, involving incrementally improved submissions, allowing banks to input the complete data set.

An alternative method, not used by GCD would be to have members submit all data and then GCD would filter out only acceptable data. This could result in biased data sets if there was any correlation between data completeness and LGD. For this reason, the more time consuming iterative process is chosen.

Similarly, GCD does not perform any cleaning or make any adjustments to data after it is accepted. When GCD reviews the received data, and compares it to other bank submissions, executives are often able to observe patterns which indicate potential errors in the data which may be systematic. For example, the use of only 0 to 1 values in a field which expects 0 to 100 values may indicate that the member has input values as decimals rather than percentages. GCD then informs the member and works with them to achieve a re-submission of affected data, rather than performing our own, potentially incorrect, fixing. GCD does not remove data from the database except with the agreement of the member who input the data. Data cannot be removed by the member banks, but can be replaced by an improved delivery. Therefore, only data which is agreed to have been incorrectly input will be removed. An example would be retail borrowers input as SME non-retail. GCD may query the member on a block of smaller borrower level exposures and limits and the member confirms that these are in fact retail data which should not have been submitted. After agreement the data would be removed. Such removals happen rarely and have only affected a small amount of data.

## OUT OF CYCLE SUBMISSIONS

New members are required to make their first submission of data outside the usual submission period in order to ensure that they receive more support from executives including a stronger data audit.

Existing members wishing to work on improving their data or extending their participation to e.g. another asset class, can request and undertake an out of cycle submission as well.

## 6.4. DATA CLEANING

As explained in section 6.2 above, all data input is cleaned by the submitting bank during the iterative process involving automated validation testing and manual data audit, including qualitative review. Therefore, every data point which is accepted into the data set is deemed clean and users of this data do not need to perform their own cleaning. They do, however, need to examine the total data pool and decide which data points are representative for them.

Readers are advised to carefully review the definitions used, calculations made and data filters applied. Evidence for the consistency and veracity of this data is presented. However, users are also cautioned that any pooled data should be carefully analysed for representativeness, before being used for benchmarking or other comparison. GCD members have access to the raw but anonymised data which enables them to produce customised representative data sets and calculate averages suitable for their own portfolio comparisons.

<sup>1</sup>For a detailed description see: B- Efron, R.J. Tibshirani, "An Introduction to the Bootstrap", Monographs on Statistics and Applied Probability 57, Chapman & Hall/Crc, 1998, Chapter 6

<sup>2</sup>B- Efron, R.J. Tibshirani, "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy", Statistical Science, Vol. 1, No.1, 54-77, 1986

<sup>3</sup>Some examples of regulatory requirements for representativeness include:

BCBS: Basel II §417, §450 and §448

European CRR: Articles 174, 179 and 185

UK PRA Internal Rating Based Approach (SS11/13) Article 10.12

EBA Guidelines on PD Estimation, LGD Estimation and Treatment of Defaulted Exposures Section 4.2.2.19

ECB's Targeted Review of Internal Models (TRIM): Section 6.2, paragraph 57 (d)(iii) and Section 2.1

US Federal Reserve: SR11-7

<sup>4</sup>The Study "A theoretical and Empirical Analysis of Alternative Discount Rate Concepts for Computing LGDs using Historical Bank Workout Data is available on the GCD Website.

## About Global Credit Data

Global Credit Data (GCD) is a non-profit association owned by 50+ member banks with the simple mission to help banks better understand and measure their credit risks through data pooling and benchmarking activities.

GCD started collecting historical loss data in 2004, to which member banks have exclusive access. This database now totals over 195,000 non-retail defaulted loan facilities from around the world.

In 2009 GCD introduced a PD database which now has over 15 years of default rates and PDs. GCD also runs a name and cluster benchmarking database to help banks calibrate and benchmark their PD, LGD and EAD models.

GCD operates all databases on a "give to get" basis, meaning that members must supply high quality data to receive data in return. The robustness of GCD's data collection infrastructure place our databases as the global standard for credit risk data pooling.

For additional information, please contact

*Nunzia Rainone*  
*Analyst & Member Support Executive*  
[nunzia.rainone@globalcreditdata.org](mailto:nunzia.rainone@globalcreditdata.org)

*Nina Brumma*  
*Head of Analytics and Research*  
[nina.brumma@globalcreditdata.org](mailto:nina.brumma@globalcreditdata.org)

[www.globalcreditdata.org](http://www.globalcreditdata.org)

