Comments in response to "A point-in-time–through-the-cycle approach to rating assignment and probability of default calibration" by M. Rubtsov and A. Petrov

Lawrence R. Forest Jr.

Aguais & Associates (AAA)

20-22 Wenlock Road, London, N1 7GU, UK

Email: LForest@AguaisAndAssociates.co.uk

Gaurav Chawla (corresponding author)

Aguais & Associates (AAA)

20-22 Wenlock Road, London, N1 7GU, UK

Email: GChawla@AguaisAndAssociates.co.uk

Scott D. Aguais

Aguais & Associates (AAA)

20-22 Wenlock Road, London, N1 7GU, UK

Email: SAguais@AguaisAndAssociates.co.uk

Received January 2017; revised April 2017; accepted 2017

Overview

In a recent paper "A point-in-time–through-the-cycle approach to rating assignment and probability of default calibration" (Rubtsov and Petrov, 2016) in this journal, Rubtsov and Petrov (henceforth "R&P") have remarks related to our work on the formulation and estimation of point-in-time (PIT) and through-the-cycle (TTC) measures of probability of default (PD). In some cases, we find that R&P describe our current work inaccurately and we clarify our approach. In other cases, we disagree with R&P and we present objections. We focus on the paper's first half (up to page 98), which is the part related to our work. As a major point, we believe that R&P should consider an alternative to the common practice of determining credit grades first, based on a scorecard scores and mappings of scores to grades, and PDs second, based on mappings of grades to PDs. We find it difficult to imagine that the ratings-first approach will explain defaults as well as one that determines PDs first and grades second, through the binning of

PDs. Under this latter, 'direct-PD' approach, one is free to explain default observations unencumbered by a prior grading model.

Detailed Comments

We define PIT (Point in Time) and TTC (Through the Cycle) PDs (and by extension PIT LGD and EAD) as follows (Aguais *et al* 2004, 2007; Forest *et al* 2013, 2015; Chawla *et al* 2015, 2016a, 2016b, 2017):

PIT PD as the unbiased, unconditional expectation of an entity's default rate over any specific horizon. A good unconditional PIT estimate should account for all relevant information including the current state of the credit cycle and unconditional outlook for its evolution. We define the one-year, TTC PD as the particular case in which credit conditions today are "normal" and evolving over the coming year as a random walk without drift.

On page 85, R&P state:

> *Without getting into a discussion of semantics, in this paper we define PIT and TTC PDs as follows. The former is an estimate of the probability of default, within the following one-year (1Y) period, based on the current state of the credit cycle. The latter represents the average 1Y probability of default during the full cycle, i.e., it is an estimate of the customer's PD, given the "normal" or long-term equilibrium state of the cycle.*

We agree with these definitions as far as they go. But we would refine them by

- clarifying that an institution needs PIT PDs for varying horizons, not only one year, and
- explaining that the TTC PD represents a way of quoting a long-run "normal" value of default distance (DD), which is a measure of instantaneous default risk.

PIT PDs at Varying Horizons. In pricing for risk, diversifying portfolios, running stress tests, and establishing loss allowances, credit institutions need to gauge risk as it is, looking ahead from the moment, and they need to evaluate it over the varying tenors of their exposures. Thus, we use "PD" in referring to the likelihood of default over many, identified horizons including one year, two years, five years, three months, eighteen months, and so on. And we include the modifier "PIT" if it's not already clear that we're referring to the best estimates of prospective risk at a given time. One may choose to emphasize the one-year PD as a reference point. But it's by no means the only PIT PD that an institution needs.

TTC PD identifies a normal DD. The Basel IRB formula calls for a TTC PD as an input, but then applies the inverse-normal, cumulative-distribution-function (CDF) to it. That converts the PD into a Probit DD. For that DD to be as required in a stress test, it must describe an entity's credit status at the moment, unaffected by any particularities of the credit-cycle outlook. The only outlook relevant to Basel RWA is the -3.09-sigma, single-factor scenario embedded within the IRB formula. Thus, if the DD arising from a PD is to depict solely an entity's current credit status, one must remove from the PD any anticipatory aspects (i.e. expected drift) and assume that credit conditions over the coming year evolve as a random walk (RW). In that case, the inverse-normal CDF produces the DD as of the moment and the related PD would amount to a quoting convention for that DD.

Further the PD/DD must be TTC. This means that the DDs for a portfolio will, when entered into the standard-normal, CDF, produce PDs that reconcile with long-run averages of realized, default rates (DRs). We say 'reconcile' rather than 'equal' to cover circumstances in which a credit portfolio has moved up or down over time in long-run-average, default risk In this case of a changed, long-run-average, risk profile, today's TTC PDs would reconcile not with the portfolio's long-run-average default rate (DR), but with that DR displaced up or down to reflect the changed risk profile.

As a last, fine point, observe that the TTC DDs in a portfolio with an unchanged, long-run risk profile will fall below the long-run averages of past DDs. The convexity of the Probit PD function within the relevant range leads to this result. The average of past, one-year PDs, reflecting a variety of one-year, credit scenarios, will exceed the PD derived (via the Probit function) from the average of DDs. Thus, to satisfy the Basel condition imposed on the RW PDs, the DD inputs will need on balance to be below long-run average, DD values. We refer to the DD values including the convexity adjustment required by the Basel condition as "normal values" in distinguishing them from long-run average values.

The magnitude of this effect depends on the amplitude of the cycle and this varies by sector. Drawing on CreditEdge EDFs, we find that, on average, reconciliation with past, average PDs involves setting the Z index to a value of about -0.35. Thus, if we assume that the RW PD at a long-run average value of Z of 0 is 1.00% and that the sensitivity of DD to a unitary Z change has a typical value of 0.20, then the TTC PD including the convexity adjustment would be about 1.21% (= $\Phi(\Phi^{-1}(.01) + 0.20 \times 0.35)$) or 1.21x the PD at a long-run average Z value of 0. The two parameters in the convexity adjustment formula arise in the construction of credit-cycle indices drawing on listed company PDs from a source such as Moody's Analytics, Bloomberg, Kamakura, or the Credit Risk Initiative of the Risk Management Institute at the National University of Singapore. We determine the sensitivity to unit Z changes (0.20 in the example) as the standard deviation of annual changes in a sector's median DDs. We determine the convexity adjustment to the long-run average Z (-0.35 in the example) as the difference between the historical average of a sector's median DDs ($= \underset{t}{\operatorname{avg}}\left(-\Phi^{-1}\left(PD_{s,t}^{median}\right)\right)$) and the DD inferred from the average of a sector's median PDs ($= -\Phi^{-1}\left(\underset{t}{\operatorname{avg}}\left(PD_{s,t}^{median}\right)\right)$).

As an aside, observe that the average value of 0.20 (= $\rho^{1/2}$) for the Z sensitivity of a firm's DD falls below values produced by the Basel correlation formula or inferred from equity returns. This reduced sensitivity is a routine result of calibrating a default model. More specifically, suppose that, similar to the approach taken with the CreditEdge model, one starts with a theoretical default distance, DD^{theo}, defined as: ln(asset value/weighted liabilities)/asset volatility. Suppose further that, within the default model, one allows that DD to be transformed by shift and scale adjustments as follows: $DD = a_0 + a_1 DD^{theo}$. In calibrating such a model using a Probit formulation, one typically gets an estimate for a_0 significantly above zero and an estimate for a_1 significantly below one (about 0.45). This second result implies that the systematic variance of the calibrated DD is only about 20% of that of the theoretical DD. As with the Black-Scholes option formula, the data force one to depart somewhat from the stylized model.

On page 87, R&P include the following comments regarding our approach.

> *While this method has a more solid basis than the variable scalar approach, it leaves a number of unanswered questions as far as we are concerned. In particular, Aguais (2008) provides few details on the underlying theoretical model. Further, it remains unclear how we should deal with the potential problem of multicollinearity, if the idiosyncratic explanatory variables in the extended PD model happen to be highly correlated with Z deviations. Additionally, Aguais says nothing about the distributional characteristics of the resulting PIT and TTC measures, which would otherwise be useful in model validations. Nor is the degree of PIT-ness formally defined, or its value estimated for an arbitrary hybrid model. Finally, the assumption of the existence of a ready-to-go customer-specific hybrid PD (obtained directly from a logistic model) as a*

> *starting point for all ensuing PD adjustments and viewing the PD master scale as a mapping from PD to rating may be seen as weaknesses of this approach. Indeed, the purpose of ratings is to use them as ordinal risk-discrimination buckets into which obligors are placed according to their characteristics. Rating-grade PDs are then estimated on the basis of the observed default rates within each bucket. Such a ratings-based approach ensures that we correct for any real-life deviations from the nice-to-have theoretical assumptions (such as linearity, smoothness, independence, normality, etc) behind whatever PD model a bank might use. If the bank obtains customer-specific PDs directly, and has a PD-to- rating master scale, then it must take care of the risk of misestimation elsewhere, while ratings become a mere communication convenience.*

Some of these remarks reflect disregard to our earlier literature (Aguais *et al* 2004, 2007) and others reflect unfamiliarity with our more recent disclosures (Forest *et al* 2013, 2015; Chawla *et al* 2015, 2016a, 2016b, 2017), one of which was a similar response article to Carlehed and Petrov (2012) article. In other cases, we disagree with the remarks.

Underlying theoretical model. With regard to the theoretical model underlying our approach, we've revealed a lot on this (Aguais *et al* 2004, 2007), but we'll repeat it here. All of our wholesale-credit models start from the Merton framework. According to that framework, default risk, broadly speaking, arises from the interplay of leverage and volatility. If leverage increases, everything else equal, default risk rises. If volatility increases, everything else the same, default risk rises. If leverage and volatility both increase, default risk rises a lot. Many empirical default models including those that we've estimated apply these ideas in a variety of ways drawing on a wide range of leverage and volatility indicators. The Z indices that we use in credit applications derive from PDs from listed-company models that embody this Merton view of default risk.

Correlation of Z indices with other PD inputs. With regard to possible correlation between Z indices and other explanatory variables in a default model, there is little that one can do about this other than await more data. Explanatory variables properly included in models are often correlated with each other and, while this may reduce the precision of the estimates, econometricians advise one to do nothing other than seek larger, estimation samples. In practice, we often find that other variables track the credit cycle to some degree and hence exhibit some correlation with the Z indices. In such cases, the Z indices in the model receive a weight below unity. Most of the time, however, we find that a legacy model tracks the cycle poorly and so the correlation of the prior inputs with the Z indices is low and the addition of the Z indices with a weight close to one improves default prediction a lot. In our experience, statistical tests routinely reject the hypothesis that the coefficient on the Z index is zero. However, there still may be a substantial margin for error, mainly due to short times series and the infrequency of recessions.

As far as correlation between a systematic risk index and an idiosyncratic indicator, that's an oxymoron.

Defining PIT-ness: With regard to 'PIT-ness,' we have defined this formally several times in each article related to PIT PDs (Aguais *et al* 2004, 2007; Forest *et al* 2013, 2015; Chawla *et al* 2015, 2016b, 2017). We estimate a legacy model's PIT-ness from the coefficient obtained on the relevant, credit-cycle indices when added as inputs to the model. Thus, if the coefficient is 0.80, we estimate that the legacy model's PIT-ness is 20%.

Previous research (Forest *et al* 2015 and Chawla *et al* 2015) provide an explicit, mathematical formulation for this conversion process. Forest *et al* 2015, page 5, equation 2.1, and Chawla *et al* 2015, page 8, equation 2.1, present a Probit-model specification repeated in (1) below.

$$
\begin{aligned}
PD_{i,t+1} &= \Phi\left(-\frac{DD_{i,t} + b_{S(i)}DDGAP_{I(i),R(i),t} + \Delta DDGAP_{I(i),R(i),t+1}}{\sqrt{1-\rho_{I(i),R(i)}}}\right) \\
DD_{i,t} &= a_{0,S(i),t} + a_{1,S(i),t}DD_{g(i)}
\end{aligned}
\qquad (1)
$$

Previous research provides definitions for all of the variables in (1). But briefly, $DD_{i,t}$ is the default distance imputed from a legacy-model's grade or PD for the i^{th} entity at time t and $DDGAP_{I(i),R(i),t}$ is the credit-cycle index at time t for a combination of the i^{th} entity's, primary industry and region. Thus, the credit-cycle indices distinguish among several, industry-region groupings. (Note that Z indices, as we define them, merely add a variance normalization to the related, DDGAP indices: Z = DDGAP/$\rho^{1/2}$.).

The coefficient $b_{S(i)}$ measures the magnitude of the needed PIT adjustment and so $1 – b_{S(i)}$ represents the PIT-ness of the legacy grades or PDs. Thus, as noted earlier, if the b coefficient has a value of 0.80 and the DDGAP index correctly measures the amplitude of the relevant, credit cycle, then, to be fully PIT, the legacy model needs an add-on representing 80% of the cycle and the legacy model itself is 20% PIT. The subscript *S(i)* indicates that this adjustment would in general vary across broad sectors or models.

Forest *et al* 2015 deals specifically with S&P and Moody's ratings. They are mapped to DDs that reflect the long-run average DRs of each of the grades. Further, those grade-to-DD mappings vary depending on whether the rating is from S&P or Moody's and whether the entity is a non-financial corporation or a financial institution. While Forest *et al* 2015 deals specifically with S&P and Moody's ratings, the legacy DDs in the formula could just as well arise from any scorecard model that produces grades and PDs. Consequently, in past work, we've applied formulas similar to this in converting a variety of legacy, PD/grading models to PIT.

PD measurement error. With regard to the distribution of estimates around true PDs, R&P correctly state that we don't dwell a lot on the margin for error in PD estimation. That same comment applies to about everyone's work on estimating PDs. The margin for error in a PD estimate for any, one entity is substantial and one can get a partial measure of that from the estimated covariance matrix for model coefficients. At a portfolio level, one can get more comprehensive measures by comparing realized with predicted DRs. However, the salient questions in choosing and evaluating a PD model include the following:

> Which of the available, conceptually plausible models best explains defaults according to the conventional metrics?
>
> Do out-of-sample tests reject the null hypothesis that the current model is the true one?

In our applied work, we routinely address those questions.

Cardinal versus ordinal, default-risk indicators. We disagree with R&P's proposition that ratings should be ordinal indicators and that our emphasis on starting with a numerical default model is a weakness.

Consider the statement that ratings should be ordinal indicators. We challenge anyone to conduct any of the basic tasks of credit-risk management on the basis of indicators of the following form: lowest risk, next-to-lowest risk, third-from-lowest, and so on up to next-to-highest, and highest. Does "lowest risk" imply a 1 bp or a 100 bps PD? Clearly one can't know based solely on an ordinal ranking and so one can't price for risk or diversify a credit portfolio properly.

Of course, R&P later introduce mappings from grades to PDs. At that point they shift to a numerical, default model, likely not the best one. Yes, in estimating a default model, one must settle on a tractable functional form. One might avoid doing that at least in the same way in starting with a grading model and then, as a subsequent step, attaching a grade-to-PD mapping. However, contrary to R&P's statement, our experience indicates that one almost always predicts defaults better by focusing on modeling defaults, unencumbered by the constraint that one must first establish a grade.

We've found that traditional grading methods, often involving subjectively determined score-to-grade assignments and grade-to-PD mappings, explain defaults less well than directly estimated default models. For one thing, human judgement seems a poor device for calibrating a model to explain rare, default events. People attach undue importance to their particular frame of reference and their own, limited experience. Reliance on limited experience adds noise. Reliance on the particularities of one's experience adds bias. Specifically, one often finds that the grades in a low-risk portfolio imply less default risk than the same grades in a high-risk portfolio. Further, one often finds that PDs arising from grading models with subjective calibrations involve a safety margin, an upward bias that a dispassionate, statistical model would avoid. Last, grading models enforce premature rounding. As we've all learned in science labs and we've confirmed in statistical trials with default models, one should resist rounding until the last step. That's what we do in mapping from PDs to grades rather than the other way around. But of course, one must accept rounding under some circumstances, because overrides, which are a part of every bank's grading process, occur only through changes to grades.

Later on, pages 88 and 89, R&P add the following description of their approach:

> *In this paper, we begin by abolishing the above assumption. In particular, we explicitly acknowledge the two-step process of obtaining a PD. Namely, we assume that the bank has a scoring function. The latter can, for example, be based on a logistic regression and produce 0-to-1 output compatible with probabilities, or be an expert based model producing integer scores, etc. Either way, the objective of the scoring function is just to order customers from low to high risk. Customers are then graded by model-specific ratings, with the latter being defined in terms of scores, not PDs. This is the first step, ie, rating assignment. The second is calibration, ie, assigning PD values to each rating grade on the basis of the observed within-grade default frequencies. Each customer is then assigned the PD of the rating grade awarded in accordance with its score. The focus of our PD adjustments is, therefore, on a rating-grade PD, not a customer-specific PD.*
>
> *A consequence of this two-step framework, and our first objective, distinguishing this paper from earlier research, is the attainment of TTC rating grades, ie, a mapping from score to rating that takes into account the cyclical movements in credit conditions.* [3] *Additionally, as touched upon in Section 1, the two-step process implies that the impact of the business cycle on PDs gets split into two parts, one responsible for rating migration, and the other for within-grade ADF variation. The concept of the degree of PIT-ness becomes more complex too. In our interpretation, this is a measure of the degree of rating migration. Hence, it is a characteristic of the rating grades, pertaining to the assignment step rather than to a particular calibration technique. Thus, we propose a new way of defining and estimating the degree of PIT-ness of a rating model.*

R&P endorse status quo: R&P's approach seems to us overly accepting of the status quo. Banks applying legacy methods commonly follow a two-step process of assigning grades on the basis of scorecard scores and then, confronted with the need for numerical PDs, appending a grade-to-PD mapping. And the idea that one can partition PIT PDs into a part explained by grades and a second part explained by a floating grade-to-PD mapping is an old one. It's intrinsic to the description of many grading models as 'hybrids,' intermediate to pure PIT or TTC models.

Indeed, we have for years started with two-step, legacy models and added Z indices to them. We call this a grade-to-PD approach. It involves minimal change to established practices. One leaves the existing processes of assigning grades and thereby PDs as they are and then:

- translates the grade-imputed PDs into DDs by applying the inverse-normal CDF,
- enters those DDs into a Probit default model, with those DD inputs potentially subject to shift and scale adjustments (i.e. DD* = a_0 + a_1 x DD(grade)),
- includes the Z indices as additional variables in the model, and
- calibrates the model to the best-available, default sample.

In our experience in cases with substantial, default samples, one gets statistically significant shift and scale adjustments and a statistically significant Z effect. Thus, through an explicit default calibration, one obtains a model superior to the one suggested by R&P.

After that, one can pursue further refinements, starting with the scores or score components or underlying, financial and judgmental inputs into a scorecard model and calibrating those more detailed inputs directly to defaults. The size of the default sample determines the extent to which, by unbundling a scorecard score, one can produce a superior default model. Over a period of time with accumulating default data, one can imagine moving from a grade-to-PD to a score-to-PD and then to a score-component-to-PD and finally to a fundamental-indicator-to-PD model. Of course, at each stage one accepts the more elaborate model only if one can reject the hypothesis that it performs no better than the prior, simpler one.

In embracing the common, two-step approach, R&P seem willing to accept the status quo without considering an alternative. We instead start with credit grading's objectives, specifically doing a good job in estimating risk-based, breakeven prices and managing the credit portfolio. We then ask how one might best address those objectives. We find that default and loss modeling answers this question. The two-step approach represents one, default-model candidate, but likely not the winning one. Observe that, coming from outside banks and credit-rating agencies, virtually all of the academics contributing to credit-risk methods focus on default and loss modeling and not on grading. R&P's approach, however, seems unwilling to challenge legacy methods.

On page 90, R&P present their method for deriving TTC grades.

> *We start by having a scoring function that produces hybrid scores (not PDs!) for customers in the development sample. The scores are then used in rating assignment. As the macroeconomic climate changes, the scores will move in a systematic manner, causing rating migrations. One way of achieving our TTC goal is to construct "floating" rating grades, i.e, grades whose boundaries will move along the cycle, so that an economy-driven change in the customer characteristics, as well as its score, will no longer result in a rating switch.*

R&P's portfolio-self-calibration proposal. To obtain such ratings, R&P suggest mapping scores to ratings on the basis of ratings, score boundaries that float in such a way as to correspond to fixed, score quantiles. As an example of this, suppose that at the start of a period, entities with scores between the values of 50 and 60 have a risk grade of 5. Suppose further that these score boundaries (of 50 and 60) represent the

60th and 70th percentile of all scores at that point in time. If scores rise during the period due to credit-cycle changes, these same quantiles will correspond to higher score values. At the end of the period, the 60th and 70th percentiles might correspond to scores of 55 and 65. These would be the score boundaries for rating 5 at the end of the period.

This way of assigning grades will lead to ratings that over time, under assumed, long-run average cyclical conditions, imply unchanging PDs under a couple of conditions. To start, the portfolio's composition in terms of TTC grades must remain the same over time. In addition, the scorecard must correctly rank the default risk of entities within the portfolio. Under these circumstances, even if scores move up and down broadly with the cycle, one may use scored quantiles in determining ratings that map to fixed, TTC PDs.

R&P emphasize that this result occurs without identifying a PD function. However, if the scores have no demonstrated relationship to PDs, we'd have no reason to use those scores and the related grades in credit-risk management. More likely, R&P have in mind at least a weakly validated, monotonic mapping from grades to PDs. In that case, the score quantiles amount to PD ones and the process is equivalent to PD-based assignments. Once again, one must explicitly consider PDs.

Further, it's easy to imagine circumstances in which R&P's approach fails to determine proper, TTC ratings. Consider, for instance, the case of an institution that decides to increase (decrease) long-run average, default risk in its credit portfolio. In that case, a score corresponding to a fixed quantile would over time represent rising (falling) rather than constant-TTC default risk.

To eliminate this possibility, R&P suggest setting the floating thresholds by chaining together estimates reflecting the static portfolio that exists at the start of each period. But this may not work either. Consider a portfolio that includes only investment-grade entities. Over a time period, eliminating cycle effects, the portfolio's TTC-risk profile would become more diffuse and some firms might cross into the sub-investment-grade range. But if one allows the TTC-score boundaries to float so that each score quantile is always assigned to the same grade as at the start of the period, the grades would remain within the initial range and the diffusion in TTC risk profile would be missed. Observe that in this case, we don't have a fixed distribution of 'true' TTC PDs, so R&P's method doesn't work.

In large dynamic portfolios in which rebalancing produces approximately a constant-TTC risk profile, R&P's approach to assigning TTC grades would work reasonably well. Dynamic portfolios more so than static ones tend to be stable in terms of TTC risk. This explains our use of large, dynamic, presumably TTC-neutral portfolios in deriving credit-cycle indices. However, in illustrating their proposed method of defining TTC grades, R&P examine static-portfolio changes.

Decomposing changes: On pages 90-93, R&P decompose variations over time in a portfolio's DR into components related, respectively, to

- migrations in grades, with each grade's DR assumed fixed,
- variations in DRs within grades, with the grade composition of the portfolio assumed fixed, and
- survivorship bias, meaning that defaults within a static portfolio cull the weaker credits.

They exclude effects of new entrants in the measurement of each, one-period change. They perform the decomposition with ratings grades represented alternatively by fixed and floating, score boundaries, with the boundaries in the second case corresponding to fixed, score percentiles.

Not surprising, when the ratings boundaries float, according to a rule designed to keep the ratings composition constant, variations in DRs within grades end up explaining almost all changes in the portfolio DR. But that merely demonstrates that, in large portfolios, systematic risk (i.e. the credit cycle) accounts for almost all of the DR variations. Idiosyncratic variations offset.

Oddly, R&P label the variations in PDs within grades a TTC part of the DR change, even though it reflects cyclical changes rather than DR changes that occur with the credit-cycle fixed at a long-run average setting. We suggest that R&P drop the TTC and PIT labels attached to the different components of the decomposition. Instead, R&P may simply refer to the three components as DR variations attributable to (1) grade migrations; (2) changes in DRs within grades; and (3) survivorship bias.

Further clarification of our approach. In response to questions raised in a review of an earlier draft of these comments, we provide answers below further clarifying our approach.

(1) *How do you convert a partly PIT PD to TTC*? We convert the partly PIT PD to TTC by removing part and not all of the current, DDGAP from the DD inferred from the PD. In the case of a partly PIT, legacy PD, the b coefficient in the formula (1) is above zero and less than one. Here we convert to TTC as follows: $DD^{TTC} = DD^{Legacy} – (1 – b) \times DDGAP$. Thus, if the legacy DD's PIT-ness is (1 - b), we must remove (1 - b) x DDGAP from it in converting to TTC.

(2) *How do you verify that the model that results from adding the credit-cycle index to a legacy model is truly PIT?* We verify the PIT-ness of a PD model by comparing the times series of average PDs that it produces with the time series of realized DRs. We routinely accomplish this by plotting the modelled PDs together with the realized DRs. The modelled, average PDs need to track the realized DRs closely with materially the same, cyclical amplitude. Only in that case would we view the model including the credit-cycle indices as PIT. As a further test, in this case of discriminatory performance, we routinely bin observations by the modelled PDs and compare average PDs with realized DRs in each of the bins. In this case, we look for that the scatter plot of average PDs and realized DRs to be tightly clustered around the diagonal line. Only in that case, do we view the model as discriminating well.

(3) *How do you ensure that in estimating a model specified as in (1) above that $0 < b < 1$*? We generally impose this condition as a constraint in the estimation. However, in rare circumstances, we may suspect that the DDGAP index understates the true cycle. In such a case, we would merely constrain b to be positive and allow $b > 1$. Alternatively, we might apply a scalar in amplifying the measured cycle.

(4) *How do you distinguish between lack of PIT-ness and effects of data lags?* We haven't focused on this other than advising modelers to try to mitigate lags in financial data by projecting financials, starting with the 'no-change from most recently reported' model and progressing beyond that only upon obtaining evidence of statistically superior performance with a more elaborate method. However, in our experience, the delays in receiving financial data account for little of a legacy model's lack of PIT-ness. Mainly, we find that the lack of PIT-ness traces to the undue stability of model inputs and this isn't fixed by shifting the timing of financial data. The market-value and market-volatility information provide the cyclical volatility that the other inputs lack.

(5) *How do you address regulatory requirements for adding regulatory margins of conservatism (MOCs) to model estimates*? In the past, we've added shift parameters to low-default-portfolio, PD models. We've determined the values of those parameters so that they raise the sample average PD by an amount implied by the assumption that the realized DR was at the 40th percentile point within the default-rate CDF.

We now believe that one should address regulatory conservatism at the broad portfolio level rather than model by model. If loss estimates on broad portfolios over several years exceed realized losses by less than a MOC determined on statistical grounds, we'd apply a scalar that increases the loss estimates by enough to reach the threshold. But, in light of the large, upward biases in today's Basel models, we don't view this as a major concern. Instead, in our current work on determining loss allowances under IFRS 9 or CECL, the salient question concerns the appropriate downward adjustments to make. Some further clarification follows.

An aggregate approach to evaluating MOCs seems warranted both by the need to reduce sampling variation in measures of estimation error and by its alignment with the capital-standard's purpose, which is to protect the solvency of the institution as a whole. A model-by-model approach, in contrast, tends to produce perverse outcomes in which risk managers that strive for highly detailed, risk quantification are disadvantaged relative to their less meticulous peers. The asymmetries in MOCs lead to this result.

More importantly, today's Basel models are producing wholesale-credit, loss estimates that are biased upwards by 60% or more, according to studies of Pillar 3 reports. Thus, the regulators would be well advised to shift emphasis from MOCs to accuracy. A heighted emphasis on accuracy would surely reduce the large, unexplained disparities in risk weights across institutions. Such disparities can't persist if models become accurate.

(6) *Are direct-PD models compliant with the Basel standards, which describe PDs as arising from grades*? In our experience, yes. We've had direct-PD models approved for use under Basel II at two banks. Also, we believe that the Basel standards were written to illustrate a common practice, not to require adherence to that practice.

In any case, our comments relate to the proper conduct of risk management. If the regulators were to prescribe the use of particular techniques, those would indeed be required for Basel RWA. Nonetheless, in its conduct of risk management, we'd advise the bank to use the best available approach, even if this differs somewhat from a prescribed, regulatory formulation. Does bank management wants to inform shareholders that it's managing risk in an inferior way?

(7) *How might a bank convince regulators that the direct-PD approach is appropriate*? We do this by making the case that we've considered all, conceptually sound approaches available to the institution and have selected the one that performs best according to the conventional metrics. The choice of models is always a comparative one – null of legacy versus alternative hypothesis of a proposed new, typically more elaborate model. We choose the latter only if one can reject the hypotheses that the greater elaboration adds no value. In our experience in conducting such comparative assessments, we often reject grade-to-PD models in comparison with direct-PD ones. Yes, internal model-validation units and the regulators will challenge models including direct-PD ones and we address those challenges as best as possible. But, in the end, the choice is unavoidably a relative one, so despite the presence of imperfections, we chose the conceptual plausible model that performs best

(8) *How do you deal with MLE estimation in the presence of default correlation*? We always include the current period, risk-factor value as an input into the estimation. This is possible, since estimation is retrospective. Assuming that that factor accounts for all correlation, the individual observations become conditionally independent. This may not be entirely true, but this represents our best effort to deal with the problem.

(9) *Can you substantiate the claim that a direct-PD model explains default better than an approach that determines grades prior to PDs*? The presumption that a direct-PD approach would perform better seems to us clear in logic. A model that optimizes the fit to defaults will in most cases explain defaults better than a model that explains defaults in some other, indirect way, particularly approaches relying on human judgment as a substitute for calibration. We can't disclose the results obtained when working for two banks. Those results, which justified the replacement of many, grade-to-PD models by direct-PD ones, is proprietary to those institutions.

We're aware of the argument sometimes made by S&P and Moody's that, by accounting for nuances in individual cases, a highly judgmental approach will perform better than a model. But our experience in cases in which the default samples allow one to test this is to the contrary. For one thing, most PD models include structured judgments as either scored or categorical inputs. Further, beyond those structured judgments, which we've found useful particularly in countering the lags in financial data, we're unsure whether the exercise of judgment really improves default prediction.

(10) *What justifies your claim that models that assign grades prior to PDs are inherently subjective?* By that claim we mean that defaults events represent the only objective (or close to objective) evidence on default risk. Thus, if grades don't derive from PDs, they must involve judgmental rather than factual assessments of default risk.

Summary

We find some of R&P's remarks on our methods inaccurate and our above comments seek to remedy those inaccuracies. In particular, we make it clear that our approach is based on the Merton framework in which default risk reflects the combined effects of leverage and volatility. We've also presented a mathematical formulation of our method for converting legacy grading/PD models to PIT and for determining the PIT-ness of those legacy models. R&P find that the floating DRs of ratings defined as score ranges representing fixed, score percentiles, explain most all the time series variation in large portfolio DRs. We point out that, so long as the underlying scores rank default risk well, the result is unsurprising. The floating DRs of the floating score bins would identify almost all of the systematic variation in default risk and the remainder, tracing to migrations with respect to floating score bins, would be unsystematic. As a last comment, we believe that R&P should consider moving away from the commonplace framework of determining credit grades first, based on scorecard scores and score-to-grade mappings, and PDs second, based on grade-to-PD mappings. We find it hard to imagine that such an approach will explain defaults as well as one that determines PDs first and grades second, through binning of PDs. In our work, we generally find that this latter, 'direct-PD' approach explains defaults better than the common grade-first method.

References

1. Aguais, S. D., Forest, L. R. Jr., Wong, E. Y. L., and Diaz-Ledezma, D. (2004). Point-in time versus through-the-cycle ratings. In The Basel Handbook: A Guide for Financial Practitioners, Ong, M. (ed). Risk Books, London.
2. Aguais, S. D., Forest, L. R. Jr., King, M., Lennon, M. C., and Lordkipanidze, B. (2007). Designing and implementing a Basel II compliant PIT–TTC ratings framework. In The Basel Handbook II: A Guide for Financial Practitioners, Ong, M.(ed).Risk Books, London.
3. Carlehed M and Petrov A, A methodology for point-in-time–through-the-cycle probability of default decomposition in risk classification systems, Journal of Risk Model Validation Volume 6/Number 3, Fall 2012 (3–25)
4. Chawla G., Forest L., and Aguais S. D., "AERB: Developing AIRB PIT-TTC PD models using External CRA Ratings", The Journal of Risk Model Validation: Volume 9/Number 4, Winter 2015, available at http://www.risk.net/journal-of-risk-model-validation/technical-paper/2437473/aerb-developing-airb-pit-ttc-pd-models-using-external-ratings
5. Chawla G., Forest L., and Aguais S. D., "Point-in-time loss-given default rates and exposures at default models for IFRS 9/CECL and stress testing", Journal of Risk Management in Financial Institutions, Volume 9 / Number 3, 2016a, pp. 249-263(15), available at http://www.ingentaconnect.com/content/hsp/jrmfi/2016/00000009/00000003/art00004
6. Chawla G., Forest L., and Aguais S. D., "Some Options for Evaluating Significant Deterioration Under IFRS9", The Journal of Risk Model Validation: Volume 10/Number 3, 2016b, available at http://www.risk.net/journal-of-risk-model-validation/technical-paper/2469333/some-options-for-evaluating-significant-deterioration-under-ifrs-9
7. Chawla G., Forest L., and Aguais S. D., "Convexity and Correlation Effects in Expected Credit Loss calculations for IFRS9/CECL and Stress Testing", Journal of Risk Management in Financial Institutions, Volume 10 / Number 1, 2017, available at http://www.ingentaconnect.com/content/hsp/jrmfi/2017/00000010/00000001/art00011
8. Forest L., Chawla G. and Aguais S. D., (2013) "Comment in response to 'A methodology for point-in-time–through-the-cycle probability of default decomposition in risk classification systems' by M. Carlehed and A. Petrov", The Journal of Risk Model Validation: Volume 7/Number 4, Winter 2013/14, available at http://www.risk.net/journal-of-risk-model-validation/technical-paper/2317266/comment-in-response-to-a-methodology-for-point-in-time-through-the-cycle-probability-of-default-decomposition-in-risk-classification-systems-by-m-carlehed-and-a-petrov
9. Forest L., Chawla G. and Aguais S. D., (2015) "Biased Benchmarks", The Journal of Risk Model Validation: Volume 9/Number 2, Summer 2015, available at http://www.risk.net/journal-of-risk-model-validation/technical-paper/2412440/biased-benchmarks
10. Rubtsov M and Petrov A, "A point-in-time–through-the-cycle approach to rating assignment and probability of default calibration", Volume 10, Number 2, June 2016, pp: 83-112