Moody's Investors Service Ltd.
One Canada Square
Canary Wharf
London E14 5FA
+44 20 7772 1000 tel
+44 20 7772 5401 fax
www.moodys.com

19 February 2016

Per Electronic Submission

European Securities and Markets Authority
103 rue de Grenelle
75007 Paris, France

COMMENTS ON EUROPEAN SECURITIES AND MARKETS AUTHORITY DISCUSSION PAPER ON THE VALIDATION AND REVIEW OF CREDIT RATING AGENCIES' METHODOLOGIES (ESMA/2015/1735)

Moody's Investors Service ("**MIS**") wishes to thank the European Securities and Markets Authority ("**ESMA**") for the opportunity to provide comments on the Discussion Paper on the validation and review of Credit Rating Agencies' ("**CRAs**") methodologies (the "**Discussion Paper**"). We appreciate the need for ESMA to continue to refine its application of Article 8(3) of the CRA Regulation.[1] Indeed, MIS continually reviews new methods of evaluating whether a methodology demonstrates predictive power, discriminatory power, and historical robustness. However, as currently constructed, we are concerned that:

- The Discussion Paper is based on a premise that credit ratings are precise, absolute and unqualified measures of credit risk. Credit ratings are forecasts, qualified by a number of assumptions about an uncertain future and rank ordered on a comparative scale.
- Considering factors such as small data samples, different time horizons and evolving credit conditions, a number of the statistical tests suggested by the Discussion Paper would add little value, produce misleading outcomes, and increase ratings volatility.

The proposed measures, when taken together, signal a level of discomfort with the innate uncertain, subjective nature of the credit ratings system. Credit rating systems are less scientific in nature than many would like to assume. A common misperception exists that credit ratings are binary – i.e., "pass-fail" or "high-low" – perhaps because bonds ultimately behave in a binary manner: that is, either they default or they do not default. At the time that MIS forms its credit opinion about any given bond, however, it is not yet known whether and to what extent the bond will perform or default. It is simply not possible to predict the future with absolute precision. For that reason, MIS has developed a non-binary rating system that reflects our view of the

[1] Regulation (EC) No 1060/2009 of the European Parliament and of the Council of 16 September 2009 on credit rating agencies (as last amended by Regulation (EU) No 462/2013).

relative future credit risk of issuers and financial obligations; and our rating scale is best understood as a relative scale, ordinal in nature.

More pointedly, no amount of validation, review and back-testing will change the fundamental essence of credit ratings or the nature of the credit rating process. ESMA has recognised that the determination of credit ratings involves subjective judgment. To that end, in assigning our credit opinions, our analysts adhere to MIS published credit rating methodologies, but bring to bear their personal perspective informed by their experience in, and understanding of, the sector. That is to say, our methodologies are not computer codes and the resulting credit ratings are not formula-driven. An assessment process that treats methodologies as codes is ill-suited for credit ratings, will provide skewed results and will inadvertently foster the misperception of credit ratings.

We would suggest that ESMA adopt an approach that requires a CRA to define the quantitative and qualitative measures used in its validation process, mapped into the requirements of the CRA Regulation and the RTS. For example, an alternative approach that fully meets the requirements of the CRA Regulation but without the issues introduced through mandatory specific statistical tests could be a process that offers:

- discriminatory power: through the evaluation of lists of defaulters and material impairments during the relevant study period, and leverages the average position (or accuracy ratio) at multiple horizons;
- predictive power: through the evaluation of default and imperative rates, as well as average cumulative default rates; and
- historical robustness through the evaluation of rating transition matrices for the most recent cohorts, historical average rating transition matrices, as well as measures of volatility and measures of reversal and drift.

The initial validation process and the remedial action required to be taken should a methodology fail the ongoing validation process should also be documented.

In the attached Annex I, we provide our general concerns with the premise of the Discussion Paper and the prominence given to statistical tests as applied to CRAs. We propose an alternative approach to achieve the regulatory objectives under the following headings: (1) absolute targets are not suitable for credit ratings on the relative scale; (2) statistical tests have limitations and can produce misleading results; and (3) a well-defined, but flexible approach is more effective for reviewing and validating methodologies. In attached Annex II, we provide responses to the specific questions raised in the Discussion Paper.

We would be pleased to discuss our views in more detail with you at your convenience.

Yours sincerely

/S/ Farisa Zarin

Managing Director and Global Head of Government and Public Affairs

ANNEX I

I. Fostering a Misperception: Absolute Targets are Not Suitable for Credit Ratings on the Relative Scale

MIS credit ratings are opinions about credit risk expressed on a relative or "ordinal" scale. This scale communicates our view of the fundamental credit risk of one entity or debt instrument *compared* to the credit risk of another similar entity or debt instrument. We attempt to group credits of similar risk profiles together by rating symbol. Obligations carrying the same rating symbol are not claimed to be of absolutely equal credit quality. Instead, each symbol represents a group in which the credit characteristics are broadly the same. The ordinal scale does not quantify credit risk in absolute terms because relative credit risk is not static and cannot be numerically measured as a unit. Rather, the ordinal scale is intended to be responsive to dynamic and evolving credit conditions over time.

Furthermore, MIS credit ratings are intended to be "stable" measures of relative credit risk, as determined by each issuer's relative fundamental creditworthiness and without reference to any one particular time horizon.[2] If the objective of a ratings system is to target specific default or loss rates by rating category, then the cardinal accuracy of a ratings system can be measured by the variability of realised default or loss rates around those targets. However, identifying "similar risk profiles" without a specific horizon on the ordinal scale requires deeper analysis.[3] More fundamentally, as ratings are opinions about probable outcomes, it is impossible to judge any single rating as "right" or "wrong" on the basis of ex-post performance.

Therefore, the adoption of a measure that seeks absolute targets would not be appropriate for MIS which does not target specific default or loss rates for its rating categories. Performance of MIS credit ratings on the ordinal scale should therefore be measured by both ratings accuracy (the correlation between ratings and defaults) and ratings stability (the frequency and magnitude of ratings changes).

II. Re-engineering Credit Ratings: Statistical Tests Have Limitations and Can Produce Misleading Results

While statistical or quantitative measures can be used as valuable tools in the evaluation of methodologies, MIS does not share the belief that using statistical *tests* as part of the validation of credit rating methodologies enhances their quality. Statistical and quantitative measures are best understood as gauges that can impart data relevant to the analysis of credit rating performance. MIS applies these measures as part of its process to review and validate its methodologies. These include default rates, downgrade rates, upgrade rates and rating transition matrices. While these tools are useful barometers of underlying credit performance, they do not speak to ratings performance by themselves. They are only meaningful when subject to qualitative analysis informed by the judgment and expertise of credit rating analysts.

---

[2] Rather than assigning a specific expected loss ("EL") or EL range to an obligation, MIS communicates through our ratings that we believe an MIS-rated Aaa obligation likely has a lower EL than one rated Aal, which we believe has a lower EL than one rated Aa2, and so on down the rating scale.

[3] Suppose two issuers are equally likely to default within the next five years, but one is much more likely to default than the other over the next year. Should they be assigned the same rating or not? No matter the answer, we cannot expect similar performance measured at both the one- and five-year horizons.

In our view, when statistical measures are applied as statistical tests, they are pushed beyond their intended purpose. Under these conditions, at best they add no value to the review and validation process, and at worst, can produce misleading results. Credit performance does not operate in isolation, but instead is impacted by a host of evolving conditions. Statistical tests do not accommodate this inherent variability, but instead rely on fixed assumptions to establish defined acceptable thresholds in order to actually determine credit rating performance. Changes in performance may be due to changes in the quality of the ratings process, or may simply reflect a change in the environment that makes defaults more or less difficult to predict. While statistical tests may provide insight into whether or not credit rating performance has changed, they do not address why it has changed. Five key limitations and concerns with the use of statistical tests as proposed in the Discussion Paper are set out below.

First, statistical tests such as the binomial and chi-square tests depend on large samples, and in some cases, on large numbers of default by rating category. These large samples are often not available as defaults are rare events. Back-testing against historical data is only feasible where adequate historical data exist, which is often not the case. Even where data exist, they may be of limited relevance where, for example, the sector has never experienced instances of significant stress that could affect the performance of highly rated securities. In such circumstances, back-testing would systematically understate the risk and lead to insufficiently conservative methodologies. For example, prior to the financial crisis, US CMBS and RMBS credit ratings were extremely stable and generally had rating performance statistics that were better than corporate benchmarks. In this instance, statistical tests alone would not suggest that additional review of the underlying methodologies might be warranted. Instead, they would suggest that CMBS and RMBS credit ratings should have been rated higher.

Conversely, it might also be possible, if unlikely, that the historical data contain a severe stress event that is no longer considered repeatable in that sector. In such a case, back-testing might again produce misleading results. A sector might, for example, have undergone sufficient structural or regulatory change over time (*e.g.*, the "bail-in" under the EU Bank Recovery and Resolution Directive[4]) so that future credit challenges would likely be different from any experienced in the past. Supplemental statistical tests, such as the Brier score and the Vasicek one-factor model test, would not correct these misleading results because they also rely on large samples and large default rates.

Second, even where data might be available for testing, the statistical tests proposed in the Discussion Paper would add little or no value to the process. For example, MIS compares accuracy ratios or average positions to evaluate discriminatory power. Adding further statistical tests to the process, including the Kolmogorov-Smirnov statistic and the ROC curve, would be redundant. The Kolmogorov-Smironov statistic seeks to determine whether two datasets differ significantly. Unless the Kolmogorov-Smironov test managed to reverse the conclusions drawn by comparing accuracy ratios or average positions, it would serve no purpose. We believe this is a highly unlikely scenario.

[4] The EU Bank Recovery and Resolution Directive ("BRRD") requires member states to enact legislation enabling regulators to modify unsecured liabilities in order to effect recapitalisation or to capitalise bridge institutions assuming the liabilities of a failing bank.

Similarly, univariate analysis of rating determinants does not add any meaningful information in the context of validating the historical robustness of credit rating methodologies. It is unclear what might be gained by studying rating determinants when the ratings themselves are of primary interest. Perhaps more importantly, credit ratings are not mechanical outputs from a model. It is unclear how the univariate analysis of rating determinants translates to meaningful inference about the ratings themselves.

Third, performance measures can be very different when measured over different time horizons. For example, during the most recent economic downturn of 2008-2009, the one-year speculative-grade corporate default reached a recent historical peak. However, the five-year speculative-grade default rate for the same cohort is about average. Based on these results, it would be difficult to draw any conclusions about the performance of the ratings based on a statistical test alone. If we calibrated to one-year performance, we might say that ratings "underperformed;" if we calibrated to five-year performance, we might say that ratings "overperformed," since default rates – given the economic stress – should have been higher.

Fourth, all statistical tests are only valid under a certain set of assumptions, many of which are not valid in the context of credit performance. The underlying assumptions are often invalid, and there are correlated linkages through the macroeconomy, financial sector, and corporate family structures such that realised default rates vary widely based on the impact of small numbers of correlated shocks.

Fifth, the proposed statistical tests are likely to place undue pressure on CRAs to adjust their credit ratings. The critical issue that a CRA considers in reviewing the performance of its credit ratings is whether to adjust forward-looking methodologies in response to past events. Any adjustment to a methodology must be forward-looking. Forward-looking considerations require a qualitative assessment of future expectations, and are outside the scope of any statistical test. The Discussion Paper's focus on statistical measures and tests, however, places the emphasis squarely on past events and prior performance. For example, depending on the statistical test applied, a short-term and isolated market event could arbitrarily induce a change to a methodology. An approach based on externally driven thresholds and triggers will undermine the deliberative credit ratings process, and may lead to unintended market disruption. Further, in addition to being inconsistent with the forward-looking nature of credit ratings, this approach is also contrary to policy efforts to inform EU investors that past performance is of "limited value as a guide to future performance".[5] In fact, the application of statistical tests and predefined reactive measures could lead to ratings volatility and pro-cyclicality.

III. A Well-Defined, but Flexible Approach is More Effective

Methodologies and credit analysis vary from one CRA to another, and the capital markets benefit from this diversity of perspectives. While MIS supports efforts to ensure that CRAs consistently apply procedures for the validation and review of credit rating methodologies, a one-size-fits-all approach is not an effective model, particularly when the model is driven by statistical tests. Rather, a transparent framework that provides for the flexible use of a combination of quantitative measures and qualitative analysis provides the best opportunity to

[5] Article 15, Section 4, Regulation (EU) No 583/2010 implementing Directive 2009/65/EC of the European Parliament and of the Council as regards key investor information and conditions to be met when providing key investor information or the prospectus in a durable medium other than paper of by means of a website.

meet these objectives, while also allowing CRAs to continue to provide diverse credit perspectives to the capital markets.

In our view, a framework for publicly available credit ratings that meets the requirements of the regulation would require:

- a process for the development of methodologies that makes justifiable use of quantitative and qualitative techniques;
- a process for the use of methodologies in the assignment of credit ratings;
- publication of default and rating performance studies; and
- a review process for methodologies that leads to changes in methodologies where appropriate.

In developing this transparent framework, the policy approach must factor in the benefits of having methodologies, ratings and performance studies that are available to the public.

These principles are largely similar to the principles proposed under the Discussion Paper, but allow for sufficient flexibility for a CRA to use the most appropriate quantitative and qualitative measures for its rating system. Mapped into the requirements of the Regulation, the principles would ensure:

- discriminatory power: through the evaluation of lists of defaulters and material impairments during the relevant study period, and leverages the average position (or accuracy ratio) at multiple horizons;[6]
- predictive power: through the evaluation of default and impairment rates,[7] as well as average cumulative default rates;[8] and
- historical robustness: through the evaluation of rating transition matrices for the most recent cohorts,[9] historical average rating transition matrices,[10] as well as measures of volatility and measures of reversal and drift.[11]

MIS evaluates the results of these exercises on a qualitative basis. This evaluation may take into account a host of considerations including, for example, the environment from which

---

6 Time series of average default/impairment position and of median rating 1 year before default/impairment for the credits studied and their relevant benchmark portfolios. The Average Position of defaults/impairments measures the ordinal power of ratings, subject to small default/impairment count limitations. The possible range for the average position is 0% to 100% with 100% indicating perfect sorting power, 50% indicating no power, and 0 indicating perfectly negative power.

7 Time series of default/impairment rates for the credits studied and their relevant benchmark portfolios.

8 Average cumulative default / impairment rates by rating at horizons of one to ten years for the credits studied and their relevant benchmark portfolios.

9 One-, three- and five-year rating transition matrices for the credits studied and their relevant benchmark portfolios. These data, while often calculated from a small sample of credits, provide insights into ratings stability for individual cohorts of credits.

10 Average one-, three- and five-year rating transition matrices for the study period for the credits studied and their relevant benchmark portfolios. This data provide insights, subject to small sample limitations, into average ratings stability for credits studied compared to the benchmark portfolios.

11 Time series of rating drift, rating volatility, reversal rates, average downgraded notches per issuer, and average upgraded notches per issuer for the credits studied and their relevant benchmark portfolios. Rating drift is defined as the average upgraded notches per issuer minus the average downgraded notches per issuer over a specified horizon. Rating volatility is defined as the average upgraded notches per issuer plus the average downgraded notches per issuer over a specified horizon.

the measures were drawn, macro-economic factors, linkages with other developments in the financial sector, and any correlations with corporate family structures. The approach, which combines quantitative factors weighted according to qualitative analysis, can only occur within a framework that is flexible and responsive to evolving credit conditions.

Finally, we also underscore the importance of affording CRAs with flexibility to determine appropriate next steps after the analysis is complete. Statistical tests cannot determine whether to adjust a forward-looking methodology based on past events. The decision to change a methodology is complex process that can only be made with qualitative analysis informed by experience and judgment. Applying statistical tests that trigger predetermined action would ultimately operate to undermine the credibility of credit ratings and the stability of the capital markets.

ANNEX II

CONSULTATION QUESTIONS

1. Do you agree with ESMA's view regarding the discriminatory power of methodologies?

Please refer to our responses to questions 2-5.

2. Do you agree that the Accuracy Ratio, as derived from the CAP curve, is the minimum statistical measure that a CRA should use as part of its validation processes for demonstrating the discriminatory power of its methodologies?

The accuracy ratio, or its equivalent the average position, adequately summarises the discriminatory power of rating methodologies.

3. Do you agree that complementary measures such as the Kolmogorov-Smirnov statistic and the ROC curve (along with a confusion matrix) add further information to the discriminatory power of methodologies? If not, please explain why.

There is very little added value from measures such as the Kolmogorov-Smirnov statistic and the ROC curve. First, although the accuracy ratio serves as a useful summary measure of a methodology's discriminatory power, it is not clear that the added measures would add further information to the discriminatory power of methodologies. Whether the goal is to compare two different rating systems, two different sectors under the same rating system or the same sector at two different times, these additional statistics would only be useful if they managed to reverse the conclusions drawn by simply comparing accuracy ratios or average positions. We believe this is an unlikely scenario.

Second, when formulated as a "test", it is not clear what the meaningful hypothesis might be. For example, one might test whether two different systems are "equally discriminatory", but only if applied on a strictly matched sample with a common definition of default. In all other cases, whether different sectors at a point in time or the same sector at different points in time, we naturally expect to see different discriminatory power. It is not clear whether the test would yield meaningful results. In fact, simply reporting the tests could incorrectly suggests that discriminatory power should always be the same.

4. Are there additional quantitative measures that CRAs should use and which would add further insight into the discriminatory power of methodologies? If yes, please explain the measures and your rationale.

Without distinguishing whether they are quantitative or qualitative in nature, three additional measures could be considered in addition to average position: (1) investment grade default rate, (2) median rating before default, and (3) default rates by rating category.

MIS routinely reports the investment-grade default rate and the median rating some months prior to default. In both cases, we expect the investment-grade default rate to be "low", and the median rating prior to default to be "low". Both measures are useful indicators of discriminatory power. Additionally, the comparison of default rates by rating category can serve as an indication of whether ratings are well-positioned on average.

5. Are there qualitative measures that are appropriate for demonstrating the discriminatory power of methodologies? If yes, please explain the measures and your rationale.

Please refer to the response to question 4.

6. Do you agree with ESMA's view regarding the predictive power of methodologies?

MIS does not agree that CRAs should define target default rates by rating category. There are a number of difficulties in attempting to do so.

First, MIS credit ratings are not intended to be cardinal summaries of targeted default rates. MIS credit rating scales are a relative or "ordinal," rather than an absolute or "cardinal" ranking system. MIS credit ratings are intended to be "stable" measures of *relative* credit risk, as determined by each issuer's relative fundamental creditworthiness and without reference to any one particular time horizon.

Second, we do not hold point-in-time expectations of default rates by rating category, nor do we expect default rates by rating category at a moment in time to be the same across all industries.

Third, it is not clear *which* default rate CRAs should target. It will generally be impossible to achieve constant performance at all horizons without focusing on one time horizon. For example, CRAs could target the one-year, the five-year, or a different timeframe. In order to focus on one time horizon, however, the CRA would be required to make the maturity date a driver of the rating. Historically, "maturity date" has not been a driver for fundamental credit ratings. Generally speaking, most CRAs rate the same class of debt coming from an issuer the same regardless of whether that debt is due to mature in the next month or in 10 years. By targeting specific default rates at a specific horizon, the CRA would need to distinguish between the two, a practice which the market is not used to and likely would find unhelpful.

Fourth, a CRA that targets absolute default rates by rating category, especially over a short horizon, will necessarily make frequent rating changes and experience frequent rating reversals, both of which are disruptive to the market and market participants. For example, during economic slowdowns, all default probabilities tend to increase, but it is not generally desired to have a CRA downgrade most of its portfolio.

Fifth, we do not see a link between realised default rates and predictive power. Rather, we view predictive power as a discriminatory exercise. To further illustrate this point, consider a portfolio of 100 credits and that it is known beforehand that 10 credits will default (although which 10 is unknown). Rating Agency Y could simply assign all 100 credits a single-B rating, resulting in a 10% single-B default rate. Rating Agency Z, on the other hand, might be able to identify 5 credits that will default with certainty, another 45 that will not default with certainty, and rate the remaining 50 credits single B. Rating Agency Z's single-B default rate will also be 10%, but their ratings are more predictive and have better discriminatory power.

If required to define expected default rates by rating category, we believe the long-term historical performance of our credit ratings are generally indicative of their future performance, provided the distribution of economic shocks experienced in the future is roughly the same as observed in the past.

7. Do you agree that statistical measures of predictive power increase the quality of validation of CRAs methodologies and should be performed by the CRAs?

While statistical measures are valuable tools in the validation of methodologies, we do not agree that introducing statistical tests as a part of the validation of rating methodologies enhances the validation process.

When validating our methodologies, MIS produces a number of statistical measures, including average position at multiple horizons, default rates by rating category at multiple horizons, measures of volatility, reversal and drift, among many others. We take note of instances where realised default rates by rating category are substantially different from benchmark sector performance measured over the same period of time. We also take notice of unusually disparate performance between sectors we would have expected to perform similarly.

While these measures are an important part of the process, they are not well-suited to serve as the basis for a formalised and mechanical statistical test. Statistical tests are valid only under a certain set of assumptions, many of which are not valid in the context of credit performance. The result of any statistical test would inevitably need to be assessed with the judgment of credit experts to determine whether the result is driven by violations of the underlying assumptions of the test. When MIS evaluates statistical measures, we apply qualitative analysis and may consider, for example, the environment from which they were drawn. This is because credit performance in a sector would rarely be construed as independent of other factors. There are correlated linkages through the macroeconomy, financial sector, and corporate family structures such that realised default rates vary widely based on the impact of small numbers of correlated shocks. For instance, we are not surprised to see high default rates during a recession. Neither are we surprised to see indications of greater discriminatory power during economic expansions.

Furthermore, supposing an instance where credit ratings underperform, a qualitative assessment of the nature of the underperformance and whether such underperformance would be expected in the future will be required. For example, suppose widespread accounting fraud in an industry leads to poor assessment of risk, and ultimately leads to several high-rated defaults. Further, suppose legal and regulatory changes follow. The critical question CRA's face is whether to adjust a methodology in response to past events. Statistical tests do not help inform that analysis because such considerations are necessarily outside the scope of any statistical test. In this instance, we would caution against adjusting the relevant methodologies since the same degree of accounting fraud is unlikely to be repeated.

8. Do you agree that the binomial and the chi-square tests are the minimum statistical measures that a CRA (when its ratings refer to default probabilities) should use as part of its validation processes for demonstrating the predictive power of its methodologies?

No. We do not agree that the binomial and the chi-square tests are the minimum statistical measures that a CRA should use as part of its validation processes for demonstrating the predictive powers of its methodologies. A number of other factors may also be useful in a CRA's validation process, including average cumulative default rates and default/impairment rates, as part of its qualitative analysis.

We also reiterate our concern with applying statistical tests to the validation of rating methodologies. In general, assumptions of independence which underlie these tests do not hold in financial and credit performance data. There are correlated linkages throughout the macroeconomy and the financial sector which typically cannot be captured by statistical validation. Also, statistical tests depend on large samples and in some cases on large numbers of defaults by rating category.

We further note that in small sectors with few or no defaults, the chi-square test does not yield meaningful results, and the binomial test is not clearly applicable to the lowest rating categories. Among those lower rating categories there may be a concern that the default rate is too low. Although a test could be formalised in that direction, the open question would be whether defaults simply have not been realised.

9. Do you agree that complementary measures such as the Brier score and the Vasicek one-factor model test add further information to the predictive power of methodologies (when the CRAs' ratings refer to default probabilities)? If not, please explain why.

In our view, measures such as the Brier score and the Vasicek on-factor model test do not add information to the predictive power of methodologies. Even assuming arguendo that they might add additional information, it is difficult to construct a scenario where these additional measures would reverse any of the limited conclusions that might be reached from use of the binomial and chi square tests.

10. Are there additional measures that CRAs should use and which would add further insight into the predictive power of methodologies when the CRAs' ratings refer to default probabilities? If yes, please explain the measures and your rationale.

As discussed in response to question 4, the investment grade default rate and the median rating before default may be useful supplementary measures to assess the discriminatory and predictive power of a ratings system.

11. Are there qualitative measures that are appropriate for demonstrating the predictive power of methodologies when the CRAs' ratings refer to default probabilities? If yes, please explain the measures and your rationale.

Please refer to the response to questions 10 and 4.

12. Do you agree that CRAs using methodologies related to creditworthiness measures other than default probabilities should use statistical measures to demonstrate the predictive power of their methodologies? If yes, please state the potential creditworthiness measures that methodologies could relate to and the corresponding statistical measures as well as any appropriate qualitative measures.

MIS credit ratings reflect both probability of default and expected loss given default (LGD). We are not aware of any case where LGD is inversely monotonic with credit ratings. Either LGD is flat, in which case all that matters is default rates, or it increases as you move down the rating scale, in which case tests on default rates are even more conservative. Further, we are not aware of any case where a rating system performs well when measured against defaults but does not perform well when measured against expected loss.

13. If ESMA establishes that there is a need for further guidance to the industry, should this guidance also cover the demonstration of predictive power of methodologies related to creditworthiness measures other than default probabilities?

No. Demonstrating the predictive and discriminatory power of expected loss rates will not yield different results from demonstrating the predictive and discriminatory power of default rates.

14. Do you agree with ESMA's view regarding the historical robustness of methodologies?

We reiterate our concern with introducing statistical tests to demonstrate the discriminatory power, predictive power, or historical robustness of rating methodologies. Rigorous statistical tests rely upon several assumptions which are not generally satisfied in the context of financial and credit performance data.

In addition, were CRAs to implement statistical tests along all three dimensions, the chances are high that if a test does not fail in one dimension, it will fail in one or both of the other two dimensions. In other words, the tests may operate at cross-purposes. For example, suppose that in response to an economic shock in a particular sector, the CRA adjusts its ratings to satisfy target default ranges. In doing so, the CRA will have compromised statistical tests on the historical robustness of its methodologies.

15. Do you agree that stability statistical measures and the transition (migration) matrices are the minimum measures that a CRA should use as part of its validation processes for demonstrating the historical robustness of its methodologies?

We believe that transition matrices may be useful in assessing the historical robustness of methodologies. We do not believe, however, that stability statistical measures such as the Population Stability Index (PSI) add any further value beyond a transition matrix. Whereas the PSI reduces overall system stability to a single number, a transition matrix allows one to understand rating stability at a more granular level.

Further, we would strongly discourage construing volatility as a rating failure *per se*. If conditions are volatile, and if ratings are to remain accurate gauges of credit risk, then ratings must necessarily also be volatile. "Spurious" or "excessive" volatility is, by definition, undesirable, but adding additional pressure to minimise volatility would interfere with the analytical process and could conflict with objectives of predictiveness.

16. Do you agree that complementary measures such as distribution analysis, the univariate analysis of rating determinants and benchmarking add further information to the historical robustness of methodologies? If not, please explain why.

We believe that analysing the evolution of a methodology's rating distribution may be a useful exercise. We do not share the view, however, that the univariate analysis of rating determinants adds any meaningful information in the context of validating the historical robustness of credit rating methodologies. First, it is unclear what might be gained by studying rating determinants when the ratings themselves are of primary interest. Second, because credit ratings are not mechanical outputs from a model, it is unclear how the univariate analysis of rating determinants translates to meaningful inference about the ratings themselves.

With respect to benchmarking ratings against external measures, it unclear how this comparison would relate to measuring the historical robustness of credit ratings. At times, it may be useful to compare credit ratings against market-derived indicators of credit risk, but not for the purpose of benchmarking. Market-derived indicators are orders of magnitude more volatile than MIS credit ratings, and if we were to benchmark against these external indicators, our ratings would be less stable. Instead, it may be useful to compare credit ratings against market-derived indicators of credit risk as a point of reference that may

prompt additional discussion and qualitative analysis. These indicators may also be taken into account when evaluating credit rating accuracy and stability, but are not appropriate for calibration of methodologies against an objective of matching the market indicators.

17. Are there additional measures (qualitative or quantitative) that CRAs should use and which would add further insight into the historical robustness of methodologies? If yes, please explain the measures and your rationale.

It may also be useful to consider upgrade rates, downgrade rates, large rating action rates, and reversal rates, as well as rating drift and volatility, to assess the historical robustness of credit ratings. While MIS reports these statistics as matters of interest, we do not generally take the position that, for example, a "high" downgrade rate is evidence of a methodology failure by itself. Rather, if further investigation indicates that the downgrades could have been avoided by a more accurate positioning of the original rating, MIS may consider changing its methodology.

18. Do you agree with ESMA's view regarding the validation of methodologies with limited quantitative evidence?

We agree that combining similar sectors and rating categories, and extending the time period over which methodologies are evaluated are all reasonable measures when validating methodologies with limited quantitative evidence. However, we do not agree that for sectors with few or no defaults it would be appropriate to consider credits that are assigned the lowest rating category as instances of "default". Doing so would be circular and not necessarily validate methodologies.

19. Do you agree that CRAs should, as a first step, investigate data enhancement in validating methodologies with limited quantitative evidence?

As noted in response to question 18, we agree that using various data enhancement techniques, such as combining similar sectors for the purposes of validation, combining rating categories, and extending the time period over which methodologies are evaluated may be reasonable measures when faced with validating methodologies with limited quantitative evidence.

However, we do not see the utility of either using third-party data to extend the data sample or creating hypothetical transactions. Both exercises are highly subjective and will not allow CRAs to demonstrate any additional rigor of their ratings. Furthermore, third-party data presents a host of challenges, including working around the assumptions and limitations of that data. It is not clear how hypothetical transactions could be simulated to represent real tests of the methodology. Such an exercise would presumably measure how well a

methodology performs against the risk characteristics used to simulate the transactions, which is only as valuable as the nature of those simulated characteristics.

20. Do you agree that CRAs should, as a second step, investigate measures that may enable them to perform statistical tests to demonstrate the discriminatory power of their methodologies?

No. Statistical tests alone are not useful in demonstrating the discriminatory powers of methodologies. This is especially true for sectors with small sample sizes where statistical tests may, in fact, produce invalid results.

21. Do you agree that historical robustness measures should be performed when validating methodologies with limited quantitative evidence?

We agree that transition matrices, rating distributions, upgrade rates, downgrade rates, large movement rates, reversal rates, rating drift, and volatility a can, on occasion, reveal methodological deficits. As previously noted, however, this is not typically the case. For example, often a downgrade rate is "high" because credit conditions worsened.

22. Do you agree that the transition (migration) matrices and benchmarking are the minimum measures that a CRA should use as part of its validation processes for methodologies with limited quantitative evidence?

Please refer to the response to question 21.

23. Do you agree that complementary historical robustness measures add further information to the validation processes for methodologies with limited quantitative evidence? If not, please explain why.

Please refer to the response to question 21.

24. Are there additional measures that CRAs should use when validating methodologies with limited quantitative evidence? If yes, please explain the measures and your rationale.

None in addition to those previously discussed.

25. Do you agree that thresholds should be set for the quantitative validation techniques?

We do not share the view that predetermined statistical thresholds should serve as the foundation of the validation and review process. The use of a threshold implies that a certain course of action is automatically required when a predefined event occurs. Arbitrary

thresholds for quantitative validation techniques would be unhelpful, and possibly detrimental, to the review and validation of credit rating methodologies.

While quantitative techniques may play a part in the validation process, performance must ultimately be evaluated qualitatively. For example, if a sector experienced significant stress, it is natural to expect elevated default rates. That is not by itself an indication that the methodology is flawed or in need of updating. In fact, it could be the case that default rates by rating category were not as high as they should have been given the nature of the stress. In fact, a mechanical evaluation against arbitrary benchmarks might indicate a "pass," while an informed qualitative assessment might indicate that the methodology has been calibrated too conservatively.

We also note that performance measures can be very different when conducted over different horizons. By way of example, during the most recent economic downturn of 2008-2009, the one-year speculative-grade corporate default reached a recent historical peak, the monthly default rate in early 2009 set a post-war record. However, the five-year speculative-grade default rate for the same cohort is just about average. Based on these results, it would be difficult to draw any conclusions about the performance of the ratings.

We are concerned that an approach that places statistical thresholds at the centre of the review and validation process would be based on the incorrect assumption that quantitative and qualitative analysis can somehow be compartmentalised as distinct and independent exercises. While the results of quantitative techniques may be informative in some instances, CRAs should not be required to act on them at face value. Instead, CRAs should be encouraged to engage in further analysis to determine the meaning of those results and the appropriate next steps.

26. Do you agree that the Internal Review Function should decide on these values?

If required to set thresholds, we agree that the Internal Review Function should decide the method by which these values are set.

27. Do you agree that predefined actions should be defined by the CRAs when the thresholds are met?

No. While we agree that a predefined process is appropriate, we do not agree that predefined actions should be laid out by CRAs when thresholds are met. First, a threshold being met for a particular measure does not necessarily suggest underperformance. Second, supposing there is in fact an instance of underperformance, the specific nature of the underperformance needs to be carefully assessed by credit experts before an appropriate remedy can be formulated. It is not possible to predefine responses which will be appropriate in each and every situation.

To the extent thresholds are set, we believe CRAs should define the nature of the additional review that may be undertaken after the application of a statistical test. For example, if a statistical test were applied and the observed default rate is determined to be unexpectedly high, an additional review of the methodology could be considered. The additional review might consist of one or more of the following elements:

- Analysis of the fundamental factors that led to the elevated rate of default to determine whether such circumstances are more or less likely to reoccur;
- Analysis of any enhancements to the methodology or its ongoing application made subsequently to the elevated rate of default to determine whether they adequately address those fundamental factors;
- Analysis of any changes to the market, regulatory or legal environment of the affected sector covered by the methodology that have occurred or are likely to occur subsequently to the elevated rate of default to determine whether they adequately address those fundamental factors.

After these and potentially other types of analysis, the methodology review could then recommend that modifications to the methodology be considered.