Re-Mapping Credit Ratings[☆]

Alexander Eisl[a], Hermann Elendner[*,b], Manuel Lingo[c]

[a] *Institute for Finance, Banking and Insurance, WU Vienna University of Economics and Business, Heiligenstädter Straße 46-48, A–1190 Vienna, Austria*
[b] *Vienna Graduate School of Finance (VGSF), Heiligenstädter Straße 46-48, A–1190 Vienna, Austria*
[c] *Oesterreichische Nationalbank, Otto Wagner Platz 3, A–1090 Vienna, Austria*

Abstract

Rating agencies report ordinal ratings in discrete classes. We question the market's implicit assumption that agencies define their classes on identical scales. To this end, we develop a non-parametric method to estimate the relation of rating scales for pairs of raters. This *scale relation* identifies for every rating class of one rater the extent to which it corresponds to any rating class of another, and hence enables a rating-class specific re-mapping of one agency's ratings to another's scale. Our method is based purely on ordinal co-ratings to obviate error-prone estimation of PDs and disputable assumptions involved, and exploits structure in the joint estimation of all rating classes' relations from a pair of raters.

We find evidence against the hypothesis of identical scales for the three major rating agencies Fitch, Moody's and Standard & Poor's, provide the relations of their rating classes and illustrate the importance of correcting for scale relations in benchmarking.

Key words: credit rating, rating agencies, rating scales, comparison of ratings
JEL: C14, G24

[☆] We are grateful to Alois Geyer, Kurt Hornik, Stefan Pichler, Amine Tarazi, and Gerhard Winkler as well as participants at the MFA 2010 conference, the SWFA 2010 meetings, and the VGSF seminar for insightful comments and discussions.

[*] Corresponding author.
Email addresses: alexander.eisl@wu.ac.at (Alexander Eisl), hermann.elendner@vgsf.ac.at (Hermann Elendner)

1. Introduction

In the aftermath of the worldwide credit crisis, credit rating agencies (CRAs) are once again in the spotlight. At least since Pinches and Singleton (1978) CRAs have been criticized[1] repeatedly: for *a*) their lack of disclosure with regard to the applied rating methodology, *b*) the potential conflict of interest, *c*) their allegedly anti-competitive and unfair business practices as well as *d*) lack of diligence and competence (see for example Frost, 2007; Teather, 2003). Especially the admittedly poor performance[2] in rating structured credit products, in combination with the oligopoly structure of the industry, has revived the discussions among researchers as well as policymakers and the general public (see for instance Hunt, 2009; Credit Rating Agency Reform Act, 2006; Lowenstein, 2008).

However, independently of the justification of criticism regarding CRAs, it is essential to also inquire about the other involved agents and thus to question the use of CRAs' assessments by various market participants who played a decisive role in exacerbating the crisis. We argue if not mis-uses at least severe misconceptions about the information provided via ratings were widespread among their users.

First, modern pricing and risk-management models require absolute levels of credit-worthiness as expressed by a probability of default (PD) as input (Kliger and Sarig, 2000). As a matter of fact a rating does not provide a PD but an assessment expressed through a rating grade (commonly a combination of letters with symbols or numbers as modifiers) and thus apparently is only an *ordinal,* forward-looking measure of an entity's creditworthiness. To obtain the absolute measure needed for pricing and risk management, many practitioners and also academics simply estimate PDs per rating class, based on the vague statements of CRAs that the likelihood of default is one of the parameters strongly influencing their assessments (Cantor and Packer, 1997). In the most basic approach, these rating-based PDs are obtained directly from annual historic default studies published by CRAs.

We object to this approach since such PD estimates are not comparable across CRAs for two major reasons. On the one hand, ratings differ across raters with respect to *a*) the underlying measure of credit-worthiness (e.g., pure probability of default estimates vs. expected-loss estimates), *b*) the time horizon (e.g., long-term vs. short-term rating), *c*) the rating philosophy (e.g., point in time vs. through the cycle) and *d*) the granularity employed in the assessment (e.g., with modifiers vs. without modifiers) (BIS, 2006; Elkhoury, 2008). On the other hand, even if all these defining characteristics were identical among CRAs, PD estimation would still suffer from portfolio effects: as the sets of

[1] Thirty-three years ago Pinches and Singleton (1978) observe: "In recent years bond rating agencies have been under increasing scrutiny because of their obvious failures to accurately predict and warn investors of impending firm-related financial difficulties."

[2] "We're very disappointed and embarassed" conceded the president of Standard & Poor's, Deven Sharma (Lippert, 2011, p. 90).

ratees differ, so do realised default rates. In the presence of contagion or other correlation (like country effects due to currency devaluations) the discrepancies can be arbitrarily large. Hence even if hypothetically rating methodologies were identical and perfectly accurate, such PD estimates cannot be compared across agencies unless focus is restricted to a common set of obligors, jointly rated by all CRAs.

Second, market participants implicitly treat rating assessments from different CRAs as equal, based on their relative order or denomination. A rating of Baa by Moody's is deemed equivalent to BBB by Standard & Poor's. This presumed equivalence of ratings is ubiquitous. Morgan (2002), as many researchers, measures disagreement of raters by split ratings, and practitioners who demand a minimum threshold of credit-worthiness define *investment grade* as a rating of at least BBB- by Standard & Poor's or Fitch, of Baa3 by Moody's or of BBB(low) by DBRS. Another example are rating triggers, i.e. clauses in loan contracts stipulating that the loan falls due in full if the company's credit rating declines below a certain level (see Atlas (2002) for their role in Enron's demise and Bhanot and Mello (2006) for a discussion when they are optimal); so is the credit-quality threshold for eligible collateral in the Eurosystem's monetary policy operations, also currently investment grade (ECB, 2008).

The question how to correctly relate rating classes across agencies becomes acute when the same entities are rated by more than one CRA. While agencies' ratings generally show a high level of agreement (Cantor and Packer, 1995)[3] there is no sound reason from which to derive equality of risk per class; the fact that rating technologies are kept secret alone casts doubt that, say, a Standard & Poor's AA rating should be exactly equivalent to Moody's Aa rating.

Failing this, however, the implications are profound: accurate measures of obligors' credit-worthiness are crucial for numerous purposes and participants, including developers of rating systems, financial institutions and supervisory authorities. As Graham and Harvey (2001) have documented and Kisgen (2006, 2009) has shown, even non-financial firms' capital structure is influenced by credit ratings.

Developers of rating systems such as banks or rating agencies need to compare ratings for at least two reasons: On the one hand, to contrast internal estimates with outcomes of other rating sources when calibrating models, in particular for segments where data are scarce (e.g., a bank relating internal estimates to rating agency data in the calibration of models for low-default portfolios). On the other hand, raters desire to contrast the qualities of competing rating prototypes. In this context benchmarking as proposed by Hui et al. (2005) proves beneficial.

Commercial banks commonly use internal ratings not only to determine the regulatory capital to be retained, but also for allocating economic capital and the pricing of obligations. Hence they need to relate ratings from their internal

[3]This is sometimes attributed to the publicity of external ratings, see for example Smith and Walter (2001).

systems to ones from external sources like rating agencies. After all, differences in risk-adjusted prices ultimately impact a financial institution's competitive position (Lingo and Winkler, 2009).

Correctly relating ratings is also valuable in the validation of rating models. Besides backtesting, benchmarking ratings from different sources is the primary aspect of quantitative validation (BIS, 2005a), particularly for models for low-default portfolios (BIS, 2005b).

The comparison of ratings across different sources needs a mapping, namely from the domain of one rater's symbols to that of another. However, the reported grades are already the result of each rater mapping their (unknown) primary estimates of credit risk via their (unkown) scale to ordinal values. Hence, for the sake of clarity we refer to the mapping of those (mapped) ordinal grades across raters as a *re*-mapping.

If it is certain that both raters' systems coincide with respect to the four defining characteristics (measure of credit-worthiness, time horizon, philosophy and granularity) as well as in their choice of scales for mapping risk estimates to ordinal grades, the re-mapping corresponds to the identity.

For instance, in the context of Basel II commercial banks are obliged to estimate standardized one-year probabilities of default if they opt for the internal-ratings-based approach to calculate capital requirements (BIS, 2006). Such a bank will employ a master scale which associates each rating class with a distinct PD interval. If then a PD estimate is available for each obligor and both master scales are known it is straightforward to (re-)map the entities across raters. In all other—and practically prevalent—cases a correct mapping is less straightforward.

Against this background our paper proposes a new, non-parametric approach to make rating assessments from different sources comparable. Our method has the advantage to allow a mapping even if the ratings differ with respect to their four defining characteristics (measure of credit-worthiness, time-horizon, philosophy and granularity), since it obviates probability of default estimation. The proposed methodology focuses on co-rated entities, i.e. obligors rated by more than one credit-assessment source. Given a sufficient number of co-ratings, we are able to relate the rating scales of different credit-assessment sources to each other. This *scale relation* enables us to compare the rating outcomes by mapping the ratings from one credit-assessment source onto the scale of another.

Based on data of all corporate long-term issuer ratings in G7 countries from the three main rating agencies Fitch, Moody's and Standard & Poor's we demonstrate how our procedure can be applied. In doing so we are able to illustrate differences among these agencies' rating behavior. We find evidence which casts doubt on the market's implicit hypothesis that equally denominated rating grades are actually equal. Furthermore, we provide functions for re-mapping the rating grades of the major rating agencies onto each others' scales. Finally, we are able to measure the unsystematic rating heterogeneity, which can be interpreted as relative rating error.

The remainder of the paper is organized as follows. Section 2 elucidates effect and importance of ratings on different scales by illustrative examples.

Section 3 lays down the formal framework we propose for the study of the relation between raters' scales. Section 4 comprises the empirical application: Co-ratings data for the three major rating agencies are employed to estimate their scale relations and the evidence against the assumption of identical scales is discussed. Section 5 concludes and points to future research.

2. Intuition

Whenever the granularity of scales differs, the need to re-map to achieve comparability of ratings becomes obvious; nevertheless, even when they do not, there remains a major leap from an equal number of classes to identical scales. To illustrate the intuition of informed comparisons, also necessary when granularity among CRAs is equal, this section analyzes the effects of systematic and unsystematic deviations in rating assessments.

The following simple example illustrates systematic deviations: Consider two experienced CRAs providing ratings on an extensive common set of obligors, both on a numerical scale from 1 to n, thus having exactly the same level of granularity. Furthermore, assume it is known that both track precisely the same measure of credit-worthiness with the same time horizon, i.e. the four defining characteristics are identical for both CRAs. Furthermore, assume it holds for all co-ratings that whenever A assigns rating i, B reports $i - 1$.[4]

One natural conclusion is that at least one of them suffers from severe and systematic bias. Treating ratings as purely ordinal information, bias is irrelevant as for any pair of obligors the same relative ranking is observed by A and B. However, since the risk measure which is estimated and mapped to the rating grades is not purely ordinal, rating bias can be important from an economic perspective. For instance, rating bias between commercial banks using the internal-ratings-based approach (IRBA) to calculate capital requirements represents a systematic deviation of one-year PD estimates, see Hornik et al. (2007b). This can bias regulatory capital and borrower selection if the bank uses the estimates also in the pricing of loans (Jankowitsch et al., 2007).

Critically though, in contrast to the banks' ratings in Hornik et al. (2007b), assessments by CRAs are public information. Moreover, given *a*) the meaninglessness of ordinal classes' numbers, *b*) how easy an agency can detect and correct for systematic bias, and *c*) the robust relation observed in an extensive sample, it follows that what appears as rating bias cannot be caused by differences in the absolute level of credit-worthiness. That level, given all other characteristics are equal, should be equal as well—and consequently, that an i reported by A simply *corresponds to* an $i-1$ on B's scale. These systematic deviation between experienced rating agencies can only be caused by a difference of their scales, i.e. the relationship between the (cardinal) risk measure and the (ordinal) rating classes.

[4] Fringe classes require more careful treatment, which we introduce in the full method next section; for this stylized illustration simply assume A never assigns 1, neither B n.

To elaborate this idea further, we extend the previous example to the case that ratings contain an absolute level of credit-worthiness and assume that the PD is known for all classes of B, and that an additional obligor is rated only by A with the same technology as the other obligors before. If A assigns, say, class 4 and we needed to produce a PD estimate, which one should we quote? One would naturally use the estimate of B's class 3. This implies an intuitive correction for the bias[5] and converts A's rating to B's scale, enabling comparability and employing the mapping that in this case yields perfect agreement.

In any practical application observed deviations are not only systematic. Instead, also unsystematic deviations are observed, and re-mappings to perfect agreement would require to violate ordinality. If any pair of obligors is not tied by A and ranked the contrary way by B, monotonic re-mappings[6] can never achieve full agreement; non-monotonic mappings, on the other hand, contradict the ordinal character of the information. Due to differences with respect to the four defining characteristics as well as to estimation errors virtually any credit-rating dataset contains such conflicting assessments.

However, with only few pairs in a large sample of high agreement,[7] should the above approach still be applied? We argue yes, especially when contrasted with the two alternatives of *a*) matching classes by their rank, which are numbers uninformative for inter-rater comparisons or *b*) using historical default rates as proxies for PD-estimates.

In essence, we argue that when relating ratings from different sources their correspondence must not be blindly assumed, but—absent theoretical determination—estimated. Moreover, a re-mapping which significantly increases agreement captures the systematic relation between two raters' classes better, and therefore the relation of their scales can be analyzed via re-mappings, by maximizing agreement.

Thereby we assume rating agencies produce unbiased estimates of credit quality. The reason is not only the simplicity of detecting and correcting for bias, but more fundamentally that biased estimates[8] are equivalent to a difference (namely a shift) in scales and vice versa. In re-mapping from one scale to another it thus is needless to account for rating bias and scale differences separately.

[5] Bias, as introduced by Hornik et al. (2007b) to the credit rating literature, and the measures of association and agreement are related to the re-mapping of ratings as discussed in detail in Appendix A.

[6] A re-mapping is monotonic if for any pair of obligors that is not tied the one rated better is never re-mapped to a class worse than the other obligor is re-mapped to.

[7] We detail in Appendix A why it is not agreement (which can be 0 even if the raters produce exactly the same risk scores, see below) but the measure of association (Kendall's τ_x) which is appropriate to judge whether a re-mapping can be meaningful (i.e. raters assess at least similar risks) because it is *scale-free*: It does not require knowledge about the scale and allows valid comparisons of raters applying different scales, even with different numbers of rating classes, thus eliminating the problem of differences with respect to granularity.

[8] To be precise, the relative bias of one rater as compared to the other is relevant. If both raters are subject to the same bias vis-à-vis some absolute scale, for instance on PDs, this common bias cannot be detected by our method since it is designed to avoid estimating creditworthiness; however, the re-mapping it yields will be as accurate as with unbiased scales.

2.1. The direct approach to re-mapping

There exists a straightforward and intuitive approach to re-mapping. While it may be inconvenient to extend and subject to some shortcomings, it is ideally suited to convey the objective of the framework presented in Section 3.

Suppose a rater A has produced ratings σ_k^A for the set of obligors which are rated σ_k^B by B, where A classifies into n^A and B into n^B classes, and $k = 1, ..., N^{A,B}$ indexes the $N^{A,B}$ co-rated obligors. Assume A wishes to transform his assessments onto B's scale. Then his task is to sort the co-rated obligors into n^B buckets based on his own original ordering. There are two possible cases concerning how much information is available to the re-rater.

In the case where A has estimated a continuous credit risk parameter s_k^A to assign his ratings $\sigma_k^A = \sigma^A(s_k^A)$, he has obtained a total order over the $N^{A,B}$ obligors via the order statistic $s_{(1)}^A > s_{(2)}^A > ... > s_{(N^{A,B})}^A$ of their credit scores. Monotonicity requires any mapping σ yielding an ordinal ranking to fulfill $s_k > s_l \Rightarrow \sigma(s_k) \geq \sigma(s_l)$. Thus A's task of mapping his metric scores onto B's ordinal scale is equivalent to placing $(n^B - 1)$ delimiters in his own ordering, defining for each obligor k the class of B he is assigned to—and thus his rating $\sigma_k^{A\to B}$ on B's scale. Note in this case it is irrelevant whether obligors are in the same class in A's reported ratings: Different scales will generally imply that some sets of obligors with the same rating by A need "breaking up", i.e. correspond to different ratings in B's terms.[9] The monotonicity requirement operates on the more informative credit scores.

In the contrasting case, an outside observer is restricted to employ no other information but the co-ratings. Consequently, (mapped) ordinal data need to be *re-mapped* to a potentially different ordinal scale, and the input is only a weak ordering. Now the monotonicity requirement becomes $\sigma_k^A > \sigma_l^A \Rightarrow \sigma_k^{A\to B} \geq \sigma_l^{A\to B}$. Consequently, when classes are broken up, it remains undefined which obligors of a given class of A should be re-mapped to higher and which to lower classes (on B's scale). While all combinatorially possible permutations are consistent with the information contained in the ordinal credit ratings, it is clear that with respect to the objective function, namely maximizing agreement, it will always be preferred to take B's assessments into account.

It is natural, then, to consider the partial ordering of obligors after *sub-sorting* the tied obligors within each of A's classes according to the assessments reported by B. This results, as depicted in Figure 1, in a single weak ordering of up to $n^A \cdot n^B$ steps[10] within which credit-worthiness of obligors is indistinguishable.

It is intuitively clear (and shown in Appendix B) that re-mapping obligors from the same sub-sorted group into different classes cannot increase agreement. Hence we can still consider re-mapping as placing $n^B - 1$ delimiters in the

[9] Mapping metric scores s_k^A onto B's ordinal scale amounts to a substitution of an appropriate mapping $\sigma^{A\to B}(\cdot)$ for $\sigma^A(\cdot)$.

[10] More precisely, this is the *maximum* number of levels in the sub-sorted ordering, since some combinations of ratings will likely be empty.

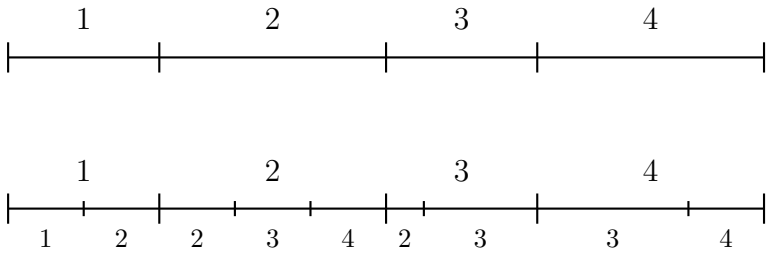

Figure 1: A's ratings, sub-sorted by B's.

ordering, now simplified to at most $n^A \cdot n^B + 1$ possible positions; often even significantly fewer as experienced raters seldom disagree by more than a few rating classes.

Effectively, this straightforward approach to re-mapping involves two steps:

1) Produce an ordering of at most $n^A \cdot n^B$ levels which is fully consistent with A's ratings, by sub-sorting within his classes according to σ_k^B.
2) There are at most $n^A \cdot n^B + 1$ boundaries of these groups at which the $n^B - 1$ delimiters of re-mapped classes can lie. For all these $\binom{n^A \cdot n^B + n^B - 1}{n^B - 1}$ cases, perform the re-mappings and find $\sigma_k^{A \to B}$ where agreement with σ_k^B is maximal.

2.2. Beyond the direct approach

The most serious drawback of the direct approach is its lack of error treatment. If at least one rater's assessments are subject to noise, this induces non-systematic disagreement. However, the direct approach provides no attempt to distinguish between the systematic relation of classes and the effect of uninformative noise. Since re-mapping is costless, a single off-by-one co-rating is sufficient to prevent the direct approach from identifying a bijective re-mapping from class a^A to the analogous a^B, irrespective of any indication of noise or identity of scales.

Not only does the direct re-mapping lend itself badly to incorporate error treatment (absent strong modeling assumptions on errors), it is moreover fairly inadequate to study the relation of two raters' scales since it produces immediately $\sigma_k^{A \to B}$ without addressing the relation of classes.

To deal with these concerns, we formalize a framework to study the scale relations directly rather than the re-mapping. We thereby aim for strict logical coherence and avoidance of both unwarranted assumptions and a concrete error modeling (which can be implemented within the framework as a next step).

The underlying intuition is again illustrated best by a thought experiment: assume two raters assess exactly the same continuous credit risk parameter,

normalized to the unit interval, for a common set of obligors. Further assume both know to estimate this credit score completely without error, thus obtaining identical scores. However not the scores but the discrete ratings are reported. Therefore they map their (identical) scores to classes denoted by integers. Now under the condition of monotonicity, i.e. ruling out worse ratings for better scores, the agencies effectively pick thresholds to delimit their classes. Assume both decide on the same number of thresholds, i.e. classes.

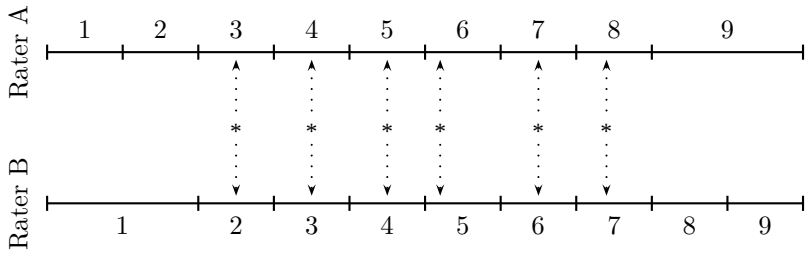

Figure 2: Disagreement purely due to different rating scales. Even identical assessments of the same risk measure will disagree if classes are defined differently.

Consider the case depicted in Figure 2, where A reports a rating of 1 for values in the interval $[0, .1)$, rating 2 in $[.1, .2)$ and the like up to rating 8 in $[.7, .8)$ and rating 9 above a score of .8, while B announces rating 1 for any score below .2, rating 2 in $[.2, .3)$, rating 3 in $[.3, .4)$ and so on up to a rating of 9 in $[.9, 1)$. For the sake of simplicity, rule out obligors with scores below .1 and above .9 to be observed. Then, clearly, all these assumptions and identical credit risk assessment notwithstanding, we find an (unweighted) agreement of 0 for these agencies' co-ratings—for no obligor do they report the same rating. The cause, however, is obviously not disagreement about credit-worthiness. The reason is (disregarding fringe classes 1 and 9) that A's rating i *corresponds to* $(i-1)$ by B.

The rest of the paper extends the logic from the last two paragraphs to the case where the risk measures may not coincide and where both raters are subject to estimation error. The aim is to disentangle the effect of (unsystematic) errors from the (systematic) relation between the respective rating classes of both.

3. A Framework for Re-Mapping

In essence, the framework addresses that what appear to be corresponding classes (like Standard & Poor's BBB+ and Baa1 from Moody's) potentially differ, since agencies might apply different rating scales. The underlying insights are twofold: First, the relation between their scales ought to be accounted for

by re-mapping to a common scale. Second, while reported ratings are ordinal, if they stem from a technology monotone in some risk measure (e.g., PD) this implies theoretical restrictions whose enforcing improves the estimation of the re-mapping. We propose the following general framework to analyze rating scales and their relation. The concrete implementation employed in the empirical part is presented in Section 3.2.

Conforming with industry practice, let every rater r produce a *credit score* s_k^r for any obligor k he assesses the credit-worthiness of. Whenever there is no risk of confusion we drop the superscript indicating the rater(s). While we allow different raters to operate different rating technologies or evaluate different risk measures, we assume s_k^r is strictly monotonically decreasing in the probability of default $PD_k^r \in (0,1]$. However, neither scores nor PDs are reported; discrete ratings on ordinal scales are.

A *rating scale* is a partition of the score range; or, due to the existence of a unique bijective link function,[11] equivalently of the unit PD range. Industry practice commonly chooses $\mathbb{R}$ as score range and the link function decreasing, so higher score corresponds to lower PD; however, we focus on the implications for the PD unit interval. Thus, a scale denotes n^r right-closed intervals $(t_{a-1}^r, t_a^r]$ for $a = 1, ..., n^r$, with $n^r - 1$ thresholds t_a^r defined (implicitly) by the rater, since $t_0^r = 0$, $t_{n^r}^r = 1$. These intervals,[12] called rating classes, are commonly labeled from AAA to D in practice, and for clarity from 1 to n^r in this paper. By the definition of partitions the intervals are non-overlapping and exhaustive.

Alternatively to intervals or thresholds, rating scales can be specified by the increasing, piecewise constant step function $a \cdot \mathbb{1}_{\{PD_k^r < t_a^r\}}$ for the unit interval as depicted in Figure 3 for two raters.

The relation between two scales is therefore fully captured by the relation of their two implied step functions. We coin the term *scale relation* to capture for every rating class a^A of rater A how much it corresponds to any class a^B of rater B, $a^A = 1, ..., n^A$, $a^B = 1, ..., n^B$. To formalize this $\{a^A\} \times \{a^B\} \mapsto [0,1]$ mapping, for the purpose of conceptual and computational convenience, define the scale relation $\varsigma^{A,B}$ as the $n^A \times n^B$ matrix $\varsigma^{A,B} = (f_{ij}^{A,B})$, where $f_{ij}^{A,B}$ denotes the fraction of A's rating class i coinciding with B's class j.

Clearly, given the rating scale of a rater A and the scale relation $\varsigma^{A,B}$ to another rater B, the latter's scale is exactly determined. Therefore knowledge of the scale relation $\varsigma^{A,B}$ enables a rater to re-map his score estimates s_k^A and convert his ratings to accord to B's definition of rating classes.

For outside observers who know no scale but only observe ordinal class ratings, a scale relation, while not sufficient to perfectly re-map on an obligor-specific level, still specifies the distribution of one rater's classes over another's;

[11] Uniqueness and bijection follow from the strict monotonicity of s_k^r in the PD.

[12] We thus adopt the convention to denote higher-PD classes with higher class numbers and speak of "better" and "worse" classes referring to those of lower versus higher PD, respectively. Also note that for rating technologies that do not produce continuous scores the link function cannot be bijective; however, a partition of the PD range still exists as long as scores are monotone in PDs.

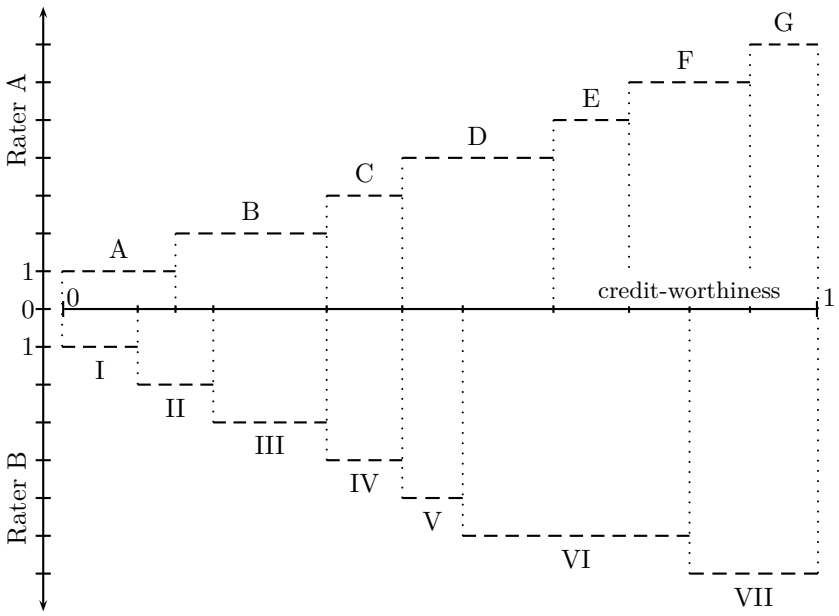

Figure 3: A rating scale is a partition of the unit interval, or equivalently a step function on the measure's range.

i.e. conveys for every class of A the fraction of ratings that correspond to various classes of B. Consequently, ignoring some residual uncertainty that we address in Section 3.2.2, scale relations define a re-mapping on the rating-class level. Such a re-rating onto a common scale is essential both to test for and to correct for the presence of different rating scales, which needs to be disentangled from rating error when calculating proximity.

3.1. Structure of scale relations

While we consciously abstain from requiring PD estimates, we conceptually anchor ratings in partitions of the PD interval to the following end. The weak modeling assumptions made so far already impose considerable structure on the scale-relation matrix $\varsigma = (f_{ij})$:

1) $f_{11} > 0$ and $f_{n^A n^B} > 0$. For $PD_k \to 0$, both raters' best classes must coincide; likewise for $PD_k \to 1$ and their worst classes.

2) $\sum_i f_{ij} = 1$ for all i. By definition of f_{ij}, all rows of ς must sum to one.

3) $(f_{ij} > 0 \wedge f_{i,j+1} = 0) \Rightarrow f_{i,j+k} = 0$ for $k \geq 1$. If a class $j+1$ of B contains only obligors of higher PD than a given class i of A, then classes of still higher PD cannot correspond to i. Analogously, $(f_{ij} > 0 \wedge f_{i,j-1} = 0) \Rightarrow f_{i,j-k} = 0$ for $k \geq 1$. In other words, every row of $\varsigma^{A,B}$ contains exactly one contiguous block of non-zero entries, conforming to the intuition that any rating class can only correspond to one continuous sequence of the other's classes.

4) $(f_{ij} > 0 \wedge f_{i+1,j} = 0) \Rightarrow f_{i+k,j} = 0$ for $k \geq 1$. If a class of A no longer corresponds to a given class of B, worse classes cannot, either. Similarly, $f_{ij} > 0 \wedge f_{i-1,j} = 0 \Rightarrow f_{i-k,j} = 0$. This is the column analogue to point 3).

5) $(f_{ij} > 0 \wedge f_{i,j+1} = 0) \Rightarrow (f_{i+1,j} > 0 \vee f_{i+1,j+1} > 0)$. Successive classes have a common boundary: If j is the last class to correspond to i, then either j or $j+1$ must correspond to $i+1$. The latter is the case if and only if the thresholds $t_i^A = t_j^B$ coincide. Additionally, $(f_{ij} > 0 \wedge f_{i,j+1} = 0) \Rightarrow f_{i+1,j-1} = 0$, i.e. B's classes of lower PD cannot match A's next class $i+1$.

Taken together, this structure requires a scale relation to resemble a connected "path" of non-zero elements through its matrix, beginning from f_{11} and ending at $f_{n^A n^B}$.

More precisely, these 5) restrictions stemming from the general framework constrict the set of permissible scale relations, and hence can and should be enforced in estimation. Table 1 illustrates the relation between the rating scales from Figure 3.

Note that the assumption of identical rating scales is equivalent to assuming all thresholds $t_a^A = t_a^B$ coincide $\forall a$, which again is equivalent to assuming the

	1	2	3	4	5	6	7
1	0.67	0.33	0.00	0.00	0.00	0.00	0.00
2	0.00	0.25	0.75	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	1.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.40	0.60	0.00
5	0.00	0.00	0.00	0.00	0.00	1.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.50	0.50
7	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 1: The scale relation of the scales depicted in Figure 3.

scale relation is the identity matrix, $\varsigma^{A,B} = \boldsymbol{I}_n$, where $n = n^A = n^B$ is the common number of rating classes. Thus, common market practice is nested within our framework, and can be evaluated like any other potential scale relation. From this point of view, we generalize existent work—where rating classes are mapped across raters by the identity matrix—to allow for all permissible scale relations, with permissibility defined by the constraints above, i.e. as conforming with PD-linked score partitions.

While identical scales clearly imply a common number of rating classes n, our framework applies as naturally with non-square matrices when $n^A \neq n^B$. In this case prior research was frequently forced to forfeit information by re-mapping the ratings with higher granularity to the coarser scale, e.g., by omitting modifiers. It is a straightforward application of our method to assess how justifiable this practice is. Moreover, it provides a means to perform this re-mapping accurately, should it still be necessary.

It is important to emphasize the mutual dependence among elements f_{ij} induced by the constraints: When comparing different scale relations, as detailed in Section 3.2.3, their matrices need to be treated as atomic, in the sense that it is in general not possible[13] to draw conclusions based on some classes irrespective of others.

3.2. Method

To address this question, we propose a non-parametric method, based exclusively on co-ratings data which works in three major steps: *a*) Construct a list of scale-relation candidates, *b*) re-map the co-ratings according to each candidate onto a common scale, and *c*) evaluate them using the proximity measure for agreement.

3.2.1. Constructing scale-relation candidates

As mentioned above, due to the strong interdependence of its elements it is in general not possible to compare potential relations incrementally. An opti-

[13] As an exception, there exists a class of dominated scale relations which can be ruled out at the outset, see Section 3.2.1.

mization over the space of permissible scale relations that does not hinge upon heuristics therefore involves exhaustively evaluating potential scale-relation candidates $\tilde{\varsigma}^{A,B} = (\tilde{f}_{ij}^{A,B})$.

To tackle the task of evaluating the infinite number of candidates, we split the problem into two simpler parts: First, construct the patterns of zero and non-zero entries in the candidate matrices; second, estimate the fractions $\tilde{f}_{ij}$ *conditional upon* those patterns.

Define formally a scale-relation (candidate) *pattern* as the binary matrix

$$\mathring{\varsigma} = (\mathring{f}_{ij}); \quad \mathring{f}_{ij} = \begin{cases} 1 & \text{if } \tilde{f}_{ij} > 0, \\ 0 & \text{if } \tilde{f}_{ij} = 0. \end{cases}$$

Given a candidate's pattern, it is straightforward to estimate the necessary fractions. This can be further simplified by excluding rating classes which only have co-ratings that fall into one category. This task thus reduces to the finite number of potential scale-relation candidate patterns.

The statement that all candidates need to be evaluated requires a minor qualification: Let an *offside pattern* denote a candidate pattern which, in some rating class, reaches far enough off the observed co-ratings so that it contains $\mathring{f}_{ij} = 1$ for at least one rating-class pair (i, j) where no co-ratings are actually observed. Then it can be proved that such patterns imply relation candidates that are dominated by non-offside patterns. For most practical applications exploiting this property by not constructing and evaluating offside patterns is crucial, since it reduces the number of candidates by orders of magnitude and renders the exhaustive search computationally feasible.

The recursive procedure we devise to construct all permissible non-offside patterns is available from the authors upon request.

3.2.2. Re-mapping ratings by scale relations

By definition, a scale relation identifies the mapping from one rater's scale to another's. Therefore, *given* the relation $\varsigma^{A,B}$ it is trivial to map scores s_k^A onto the scale of B, and rating agencies can employ our framework to accurately assess their ratings on other agencies' scales, e.g., for correct comparison.

However, since raters commonly neither report their scores nor scales, observers are constrained to re-map ordinal ratings that have already been mapped via a scale. Clearly in the first mapping, which reduces to ordinal data, information is lost; consequently no re-mapping can reproduce a direct mapping of scores via ς exactly. This limitation notwithstanding, a scale relation is by construction well suited to re-map ordinal ratings. Each row can be interpreted as a conditional re-mapping rule: Consider all obligors rated into class i by A: then the fraction f_{ij} belongs to class j in B's terms, and should be re-rated there for a correct comparison on an identical scale (in this case, B's).

This defines the re-mapping for the aggregate class, while it does not tie down on an obligor-specific level who should belong to which fraction—this is the information lost with the first mapping to ordinal. Therefore, for fixed i,

the decision problem arises to separate i-rated obligors into as many ordered subclasses as there are fractions f_{ij} which the relation specifies as non-zero. Note that it is not necessary to produce an ordering of all those obligors, it suffices to separate them into (mostly very few) lower-PD to higher-PD subsets.

Three basic approaches can be taken:

1) Rely only on A's published information; accordingly acknowledge that obligors cannot be distinguished and sample randomly. In this case it is imperative to integrate out sampling effects.
2) Also consider B's published information; thus conclude that given an identical rating by A, ceteris paribus obligors with higher rating from B should be classified into better classes.
3) Incorporate information by other raters, possibly excluding B. If co-ratings of the obligors exist with other raters, it appears natural to exploit this information. However, when more than one rater is taken as reference, the question arises how to construct their consensus opinion, which is beyond the scope of this paper.

To motivate our approach, we consider the (hypothetical) case where we know the rating scales of both A and B, so we can calculate their true scale relation analytically. Assume furthermore that both assess true scores perfectly accurately. Due to the absence of rating errors, all disagreement stems from their different rating scales. Under these circumstances, as a design goal we would want our algorithm to reproduce the analytical scale relation. It is easy to show that this requires approach 2) to re-rating.

In the presence of rating errors, the re-rating procedure suffers in proportion to their magnitude: The higher the impact of noise on estimated scores, the more the approach will resemble 1).

It is key to correct apprehension of the proposed method to differentiate between the rating information provided by B per se (on his scale), and its usage in the conditional re-rating. Only the information from the partial ordering of B is employed in the re-mapping, which moreover is carried out *conditional* both upon the ratings of A and the scale relation.

To illustrate, assume A assigns 100 obligors to his class 5, and the scale relation $\varsigma^{A,B}$ indicates this class corresponds to B's classes 3 and 4 in equal proportion. Then B's information will only be used to determine which 50 obligors should be re-mapped into the better class and which 50 into the worse. If, for the sake of the argument, B classifies 50 of those obligors as (his) class 7, another 30 as class 8, and 20 as class 9, this does not imply that anyone is re-rated to classes 7–9. Given the scale relation, the first 50 are considered as rated 3 (on B's scale) by A, while those in classes 8 and 9 (as reported by B) to be rated 4 by A, again on B's scale.

This illustration also makes clear that a re-mapping can easily decrease agreement, and in particular elucidates that a scale relation too accommodation to one rating class, by virtue of the structure embedded in our framework, impairs agreement through its need to impose harsh restrictions on the other classes.

3.2.3. Evaluating scale relations

An exhaustive search over scale-relation candidates begs the question how to judge them against each other and the identity matrix benchmark. We need to capture the degree to which two raters assign obligors the same rating class. This is exactly what is captured by the agreement measure κ, which is briefly re-examined in Appendix A.

4. Empirical Analysis

We compile a dataset of long-term corporate issuer ratings from the three major rating agencies Fitch, Moody's and Standard & Poor's. Ratings for all companies headquarted in a G7 country are obtained from Reuters Credit Views as from February 16$^{\text{th}}$, 2009. In total, 2721 obligors are co-rated, 403 thereof by all three agencies; the exact number of co-ratings per pair is given in Table 2.

	SnP	Moody	Fitch
SnP	2672		
Moody	616	665	
Fitch	2459	452	2508

Table 2: Number of co-ratings in the G7 dataset.

Our framework is static in the sense that it estimates the scale relation at one point in time. Since the systematic relation should be fairly stable over time, high volatility of ς in the time series would cast doubt on the estimation procedure. Moreover, if agencies do not adjust ratings simultaneously, this could contaminate the estimates. To safeguard against these objections, we re-run our analyses on different earlier dates and find the results qualitatively unchanged.

While the rating agencies abstain from an exact definition of the credit-worthiness they estimate and keep their rating models secret, it is clear they assess at least highly related concepts. This can be seen in Figure 4, a scatterplot of the co-ratings from Moody's and Standard & Poor's,[14] where the reported ratings are plotted with jittering, i.e. an added random disturbance to make the amount of individual data points visible on the discrete grid. Dotted boxes indicate "corresponding" rating classes, grouping those which differ only by modifier.

Although the plot is guilty of one of the main points this paper critises, namely implicitly treating ordinal rating data as if it were metric (and assuming equi-distant adjacent classes), we include it to illustrate two important facts: First, unclarity about their construction notwithstanding, agencies' credit-risk

[14] Scatterplots of the other pairs of raters are similar and thus available upon request.

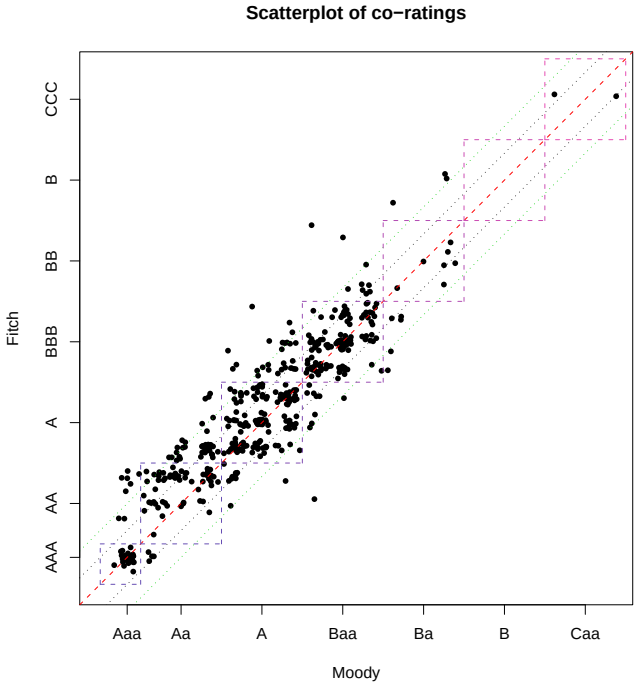

Figure 4: Scatterplot of Moody's and Fitch with jittering.

measures are indeed highly related. We quantify this prerequisite to meaningful scale relations appropriately (i.e. with ordinal statistics) in Section 4.1.

Second, the graph is suggestive of asymmetry and class-dependent variation. For instance, note that while a reasonable number of Aaa-rated obligors exhibit ratings from Standard & Poor'sof AA- not a single AAA-rated entity received a Moody's rating worse than Aa+. Also, the variation around AA/Aa appears discernibly lower than around e.g., BBB+/Baa+. Both effects would require a complex structure on the measurement error if rating classes corresponded exactly, while they arise naturally with non-identity scale relations.

4.1. Proximity measures

The following tables show the proximity measures calculated for the three rating agencies and include bootstrapped standard errors. Although the measures appear rather high, their standard errors are small, making differences between the rating agencies easily significant. This also indicates that improvements due to re-mapping need not appear high in absolute terms in order to constitute a significant improvement. Table 3 shows agreement, Table 4 association, and Table 5 bias as defined in Appendix A.

Table 3: Agreement

	SnP	Fitch	Moody
SnP	1.0000000		
Fitch	0.9690730	1.0000000	
	(0.002031639)		
Moody	0.9054569	0.9054834	1.0000000
	(0.009074744)	(0.009573705)	

Table 4: Association

	SnP	Fitch	Moody
SnP	1.0000000		
Fitch	0.8744840	1.0000000	
	(0.00618836)		
Moody	0.7524240	0.7555853	1.0000000
	(0.01268925)	(0.01341078)	

Table 5: Bias

	SnP	Fitch	Moody
SnP	0.000000000		
Fitch	-0.005040139	0.000000000	
	(0.0008120209)		
Moody	-0.022495446	-0.021358025	0.000000000
	(0.0027663372)	(0.0027916832)	

4.2. Results

		Moody						
		Aaa	Aa	A	Baa	Ba	B	Caa
Fitch	AAA	1	0	0	0	0	0	0
	AA	0.09	0.91	0	0	0	0	0
	A	0	0.196	0.804	0	0	0	0
	BBB	0	0	0.172	0.828	0	0	0
	BB	0	0	0	0.562	0.438	0	0
	B	0	0	0	0	1	0	0
	CCC	0	0	0	0	0	0	1

Table 6: Optimal scale relation on empirical co-ratings for Fitch and Moody's.

Table 6 shows the results from estimating the optimal scale relation according to our framework for Fitch and Moody's.[15] We can observe that at least for classes of lower credit risk (i.e. lower-number classes) there is indication of a shift between the scales of Moody's and Fitch, although the fractions of classes re-mapped to classes of different rank remain modest to small.

For the higher-risk classes the dataset is thinner, as was visible from Figure 4, and estimation becomes harder and more imprecise (which we show in the next section). Since some fractions are small enough to be rounded to zero, we indicate the path of the optimal scale relation by gray background behind the numbers.

However, the depicted values depend on the realized sample. If we acknowledge that co-ratings contain a stochastic element due to noise in their estimation by agencies, the observed sample is one draw from the underlying (joint) distribution. This raises the question how far the scale relation found so far is subject to sample effects, or, in different words, robust to re-sampling from the empirical distribution of co-ratings. To assess this crucial issue without an ad-hoc specification of the error structure we employ a bootstrapping procedure.

4.3. Bootstrapping

As mentioned, the scale relations estimated above depend on the samples at our disposal. In order to account for noise in the ratings, we bootstrap the cross-tables 1000 times each, i.e. draw 1000 same-size samples of co-ratings with replacement from their empirical distribution. Note that this approach does not require modeling the distribution of the errors but only independence between co-ratings of different obligors.

To gauge the sensitivity of the optimal scale relation to different draws from the distribution of co-ratings, we report in each element in Table 7 the percentage of times (from the 1000 samples) that a scale relation passed through that

[15] Results on scale relations between the other pairs of raters are similar and will be included in a future revision of this paper.

		Moody						
		Aaa	Aa	A	Baa	Ba	B	Caa
Fitch	AAA	**1*****	0.156	0	0	0	0	0
	AA	0.844	**1*****	0	0	0	0	0
	A	0	**1*****	**1*****	0.006	0	0	0
	BBB	0	0	**0.994*****	**1*****	0.176	0	0
	BB	0	0	0	0.824	**1*****	0.042	0
	B	0	0	0	0	**0.958****	0.042	0
	CCC	0	0	0	0	**0.958****	**1*****	**1*****

Table 7: Fraction how often the optimal scale relation linked any two rating classes in 1000 bootstrapped crosstables for Fitch and Moody's.

position for Fitch and Moody's. Note that these percentages are based solely on the binary information in order to pinpoint the variability in the *structure* of the relation. Positions with entries close to one are contained in virtually all scale relations: The (at least partial) correspondence of the respective classes are robust to the variation in the co-ratings. If such elements are off the main diagonal this casts doubt on the one-to-one mapping commonly assumed. Table 8 and Table 9 show the same results for Moody's and Standard & Poor's, as well as Fitch and Standard & Poor's.

		SnP						
		AAA	AA	A	BBB	BB	B	CCC
Moody	Aaa	**1*****	**0.999*****	0	0	0	0	0
	Aa	0.001	**1*****	**1*****	0	0	0	0
	A	0	0	**1*****	**1*****	0	0	0
	Baa	0	0	0	**1*****	**0.959****	0	0
	Ba	0	0	0	0.041	**1*****	0.4	0.043
	B	0	0	0	0	0.6	0.357	0.043
	Caa	0	0	0	0	0.6	**0.957****	**1*****

Table 8: Fraction how often the optimal scale relation linked any two rating classes in 1000 bootstrapped crosstables for Moody's and SnP.

The results of the bootstrapping procedure elucidate that ideally the information obtained via the re-sampling procedure should be incorporated in the estimation of the scale relations. Otherwise the scale relation is estimated by just one optimization of agreement, and thus conditional on the sampling distribution. By using the information obtained by bootstrapping this conditionality can be addressed and a potential bias in the fractions of scale relations prevented; this issue is detailed in future revisions of our paper.

In any case it is important to highlight that for all three pairs of rating agencies we find that several positions off the main diagonal have values of 1,

		SnP						
		AAA	AA	A	BBB	BB	B	CCC
Fitch	AAA	**1*****	0.87	0	0	0	0	0
	AA	0.13	**1*****	**0.999*****	0	0	0	0
	A	0	0.001	**1*****	**0.999*****	0	0	0
	BBB	0	0	0.001	**1*****	0.859	0	0
	BB	0	0	0	0.141	**1*****	0.383	0
	B	0	0	0	0	0.617	**1*****	0.884
	CCC	0	0	0	0	0	0.116	**1*****

Table 9: Fraction how often the optimal scale relation linked any two rating classes in 1000 bootstrapped crosstables for Fitch and SnP.

indicating clearly a systematic relation between rating classes that differs from identical rating scales. While the thresholds defining the classes on the scales of Standard & Poor's with respect to Fitch as well as to Moody's appear shifted, the relation of the scales of Moody's to Standard & Poor's seem to exhibit a more complicated pattern, with some classes specified broader by one and some classes broader by the other agency.

5. Conclusion

Rating agencies report ordinal ratings in discrete classes. We question the market's implicit assumption that agencies define their classes on identical scales. To this end, we develop a non-parametric method to estimate the relation of rating scales for pairs of raters. This *scale relation* identifies for every rating class of one rater the extent to which it corresponds to any rating class of another, and hence enables a rating-class specific re-mapping of one agency's ratings to another's scale.

In its simplest application, the re-mapping based on an estimated scale relation is equivalent to a straightforward direct re-mapping: Produce a weak ordering by sub-sorting your rating assessments according to the information provided by another rater, then subdivide this ordering into as many classes as the competitor's scale encompasses. By maximizing agreement the re-mapping is aligned with the external scale. However, this re-mapping is conditional upon the concrete realization of co-ratings and thus treats effects from random noise deficiently.

In the presented framework for the estimation of scale relations it is easily possible to specify a desired error modeling; in addition, it is also possible to draw inference without doing so by bootstrapping from the empirical distribution of co-ratings. In this way, we find that for the three major rating agencies Fitch, Moody's and Standard & Poor's the deviations from identical scales of long-term corporate issuer ratings of corporations in G7 countries are too pronounced to be attributed to random chance.

We thus conclude that the implicit assumption of identical scales and hence the common regulatory and industry practice of equating rating outcomes from these agencies seems doubtful. Due to the critical importance for financial institutions, rating agencies, and supervisors to accurately assess credit risk, as exemplified also in the current credit crisis, this topic calls for further research.

A. Proximity measures and their relation to re-mapping

A.1. Proximity measures

As rating data is not metric but ordinal, consistency requires to rely on appropriate measures for their relation. The application of contemporaneous versions of the proximity measures Cohen's κ and Kendall's τ (as well as the introduction of a measure for bias) was pioneered by Hornik et al. (2007b); empirical results based on these measures can be found in Hornik et al. (2010) and Hornik et al. (2007a). In the following we give a brief recollection of the proximity measures used.

All measures are defined for pairs of raters A and B, where the calculations build on those of their ratings σ_k^A and σ_k^B for which the obligors $k = 1, ..., N^{A,B}$ are (co-)rated by both A and B.

A.1.1. Agreement

Agreement captures the degree to which two raters assign obligors into the same rating class. Note that this only makes sense if the raters assign into a common number of rating classes n.[16] The classical measure, Cohen's κ, quantifies if agreement is better than ($\kappa > 0$), equal to ($\kappa = 0$), or worse ($\kappa < 0$) than by chance. The intuition builds on a cross-tabulation of ratings: Let the matrix $\boldsymbol{C}^{A,B} = (p_{ij}^{A,B})$ tabulate the observed relative frequency of obligors rated as class i by A and j by B, so

$$p_{ij}^{A,B} = \frac{\#\{\sigma_k^A = i,\ \sigma_k^B = j\}}{N^{A,B}}.$$

Ratings on the main diagonal of $\boldsymbol{C}^{A,B}$ are clearly in agreement. However, considering only them as agreeing would treat an obligor rated differently by only one notch the same as one rated AAA by one agency and C by the other. Thus the literature has often considered the first case fractional agreement rather than complete disagreement. This implies, instead of weighing the main diagonal with 1 and the rest with 0 as Cohen's κ does, weights that are a decreasing function of the difference in rating classes, gradually falling from 1 on the main diagonal to complete disagreement furthest from it. One common choice makes the function linear; in the credit-rating literature Hornik et al. (2007b) suggest the weights proposed by Fleiss and Cohen (1973), quadratic in the difference, $\boldsymbol{w} = (w_{ij})$ with $w_{ij} = 1 - \left(\frac{i-j}{n-1}\right)^2$.

The agreement-weighted sum of observed relative frequencies of co-ratings $P_o^{A,B}$,

$$P_o^{A,B}(\boldsymbol{w}) = \boldsymbol{w} : \boldsymbol{C}^{A,B} = \sum_i^n \sum_j^n w_{ij} p_{ij}^{A,B}, \tag{1}$$

however, does not consider that—even with independent ratings—some co-ratings are expected to lie on or close to the main diagonal. Assuming the

[16] All matrices in the calculation of agreement have dimension $n \times n$.

agencies rated independently, the expected cross-table is $\boldsymbol{C}_x^{A,B} = (p_{i\cdot}p_{\cdot j})$, comprising the products of the marginal proportions $p_{\cdot j} = \sum_i^n p_{ij}$ and $p_{i\cdot} = \sum_j^n p_{ij}$. Consequently, for given weights,

$$P_e^{A,B}(\boldsymbol{w}) = \boldsymbol{w} : \boldsymbol{C}_x^{A,B} = \sum_i^n \sum_j^n w_{ij} p_{i\cdot} p_{\cdot j}.$$

(Weighted) κ subtracts this correction before normalizing maximal agreement to 1:

$$\kappa^{A,B}(\boldsymbol{w}) = \frac{P_o^{A,B}(\boldsymbol{w}) - P_e^{A,B}(\boldsymbol{w})}{1 - P_e^{A,B}(\boldsymbol{w})}$$

Finally, it is important to note that the choice of a weight matrix implicitly treats the data as if they were on an interval scale since it specifies relative distances between classes. Only the unweighted case (where $\boldsymbol{w}$ is the identity matrix) is theoretically fully consistent with ordinal—even nominal—data.

A.1.2. Association

We measure association (also denoted rank correlation) with τ_x, the e*x*tension of Kendall's τ developed by Emond and Mason (2002) as the unique rank correlation coefficient to fulfill the elementary axioms outlined by Kemeny and Snell (1962). It differs from τ only in the treatment of ties, yet ensures any ranking to be perfectly correlated with itself,[17] and the triangle inequality to hold, i.e. the distance of two objects cannot exceed the sum of their distances to a third one.

Association quantifies the extent to which two agencies report the same *relative ordering* of obligors. The ranking of rater A is condensed in the $N^{A,B} \times N^{A,B}$ score matrix $\mathbf{A} = (a_{ij})$ where[18]

$$a_{ij} = \begin{cases} 1 & \text{if obligor } i \text{ is ranked ahead of or tied with obligor } j, \\ -1 & \text{if obligor } i \text{ is ranked behind obligor } j, \\ 0 & \text{if } i = j. \end{cases}$$

Given the analogous definition of $\mathbf{B} = (b_{ij})$, similarity with regard to a single pair of obligors is indicated by $a_{ij}b_{ij} > 0$, while dissimilarity entails a negative product. The measure $\tau_x^{A,B}$ is then defined as the sum of these products scaled to the $[-1, 1]$ interval; or, equivalently, as the Frobenius inner product of the score matrices divided by its maximum possible value:

$$\tau_x^{A,B} = \frac{\mathbf{A} : \mathbf{B}}{N(N-1)} = \frac{\sum_{i=1}^n \sum_{j=1}^n a_{ij} b_{ij}}{N(N-1)}$$

[17] In fact, Kendall's τ is undefined for all-ties rankings.

[18] The requirement that any ranking be perfectly correlated with itself implies that a_{ij} needs to be defined as ± 1 when i and j are tied. Only then will $a_{ij}^2 = 1$ for all $i \neq j$ and $\tau_x^{A,A} = 1$ for any $\boldsymbol{x}^A$.

The intuition of the measure is to consider the $N^2 - N$ pairs[19] formed from the co-rated obligors and compare the assessments of the two raters: Those pairs where one rater ranks the first obligor below the other while the second rater disagrees, decrease $\tau_x^{A,B}$; the other pairs, where the raters are considered in accord, increase it. The denominator ensures τ_x maps to the interval $[-1, 1]$, where $+1$ indicates perfect association, i.e. all the pairwise ratings are in the same order, and -1 indicates the opposite.

It is important to note that τ_x is *scale-free*: Because only relative orderings are considered, no knowledge about the scale is required and valid comparisons of raters with different scales, even with different numbers of rating classes, can be drawn.

Finally, in the context of credit ratings high τ_x indicates the two agencies assess identical or highly correlated risk measures. With a minor qualification[20] this is independent of the scale they employ (and in particular independent of using the same scale), and thus necessary for any re-mapping to make sense.

A.1.3. Rating bias

The average number of rating classes which A's assessment lies above B's, scaled to the interval $[-1, 1]$, was defined as *rating bias*:

$$\theta^{A,B} = \sum_{i}^{n} \sum_{j}^{n} \frac{i-j}{n-1} p_{ij}$$

Equivalently, and likely simpler to calculate, $\theta^{A,B}$ equals the difference of the mean ratings, divided by $n - 1$:

$$\theta^{A,B} = \frac{\bar{x}^{AB} - \bar{x}^{BA}}{n-1}$$

Yet another way to obtain $\theta^{A,B}$ is as the intercept in a linear regression of $\boldsymbol{x}^{AB}$ on $\boldsymbol{x}^{BA}$, with the slope parameter restricted to 1.[21] This perspective allows, provided correct handling of standard errors, an analytical significance test of rating bias.

A.2. Lack-of-proximity patterns and their interpretation

Lack of proximity arises in three cases: *a*) Raters estimate different risk characteristics (eg., PD vs. EL), *b*) they map rating scores differently to rating classes, or *c*) at least one of them estimates inaccurately.

[19] The "pairs" of obligors with themselves are excluded as they are trivially always tied and thus contribute no information.

[20] Because of the treatment of ties even independent random ratings have no expected τ_x of zero: it depends on the number of ties, and thus on the number of rating classes. Therefore a potential discretization effect arises when continuous data (like a rating score or PD estimate) are discretized (to rating classes) insofar as estimates close to the discretization thresholds can be mapped to classes with different numbers of ties.

[21] Since the slope is restricted to 1, it is equivalent to regress $\boldsymbol{x}^{BA}$ on $\boldsymbol{x}^{AB}$.

Case a), with the risk characteristics closely correlated, can be treated econometrically as if different scales were employed; the proposed method then yields the mapping from one risk measure to the other. If the characteristics have only a weak relation,[22] the case is indistinguishable from c).

Case b), the focus of this paper, is indicated by high τ_x, low κ and potentially high θ. Since the association measure does not require a common scale, τ_x can be validly calculated directly on the raw rating data as provided by both raters. If high, by definition of τ_x, the raters tend to judge the same obligor in any co-rated pair less risky,[23] which indicates that raters estimate the same (or highly correlated) risk. After re-mapping, which puts the ratings on a common scale, θ should get close to 0 and κ significantly higher.

Case c) implies both τ_x and κ are low (with arbitrary θ), and additionally that no re-rating by a different scale will be able to significantly improve κ.

A.3. First step of re-rating: Correcting for bias

A first step to re-map the data to a common scale is to correct for θ. Since θ is by definition the systematic difference in two sets of ratings, a shift of $\alpha = \theta(n-1)$ removes the bias. If $\alpha = 1$, as in the example in Section 2, treating all ratings of B as one class higher allows for an unbiased benchmarking of the two raters. However, given the discrete nature of rating classes, a non-integer shift of ratings is not meaningful. The appropriate interpretation of a fractional α of, say, .5 in this context is not that all ratings are on average rated half a class higher, but that, on average, ratings are one class higher for half of the obligors. Therefore we argue that to correct for any bias a first step shifts ratings by the integer part of α; the second step re-maps the number of obligors per rating class given by the fraction of α. The question whom to select into this percentage and how to employ co-rater information is covered in Section 2.1.

While treatment of the bias illustrates the underlying insight into the necessity of re-mapping clearly, it is evident that the elementary correction stated in the last paragraph implicitly imposes severe restrictions on the structure of the bias which remains doubtful.[24] The framework in Section 3 can be viewed as per-class specific generalization of the bias correction.

B. Sketch of proof: Identically sub-sorted groups are re-mapped jointly

Assume that we have a cross-tabulation of co-ratings $\boldsymbol{C} = (p_{ij})$ such that there exist at least two co-ratings in an off-diagonal element $p_{ij} > 1, i \neq j$. Now

[22] Negative correlation appears unlikely since raters commonly do not estimate sufficiently different concepts of credit risk. The case would be indicated by $\tau_x \ll 0$ and could be dealt with by our method slightly adapted.

[23] Note that nevertheless even identical scores generally produce $\tau_x < 1$ when mapped to different rating scales. Scores near the boundaries are mapped into different classes and consequently tie with different obligors. For instance, PDs of $(.01, .02, .03)$ can be mapped by different scales to $x^{AB} = (1, 1, 2)$ and $x^{BA} = (1, 2, 2)$ and thus give $\tau_x^{A,B} = \frac{1}{3}$.

[24] While it seemingly imposes a uniform distribution over rating classes, disputable in itself, this is hard to reconcile with any (yet unspecified) treatment of fringe classes.

it follows from the definition of the weights as non-increasing in Appendix A.1.1—which follows from the transitivity property of ordinal data—that re-mapping an off-diagonal element closer to the main diagonal cannot decrease κ. Thus there remain two cases:

1) In the corner case where the re-mapping does not affect agreement, all p_{ij} obligors can be re-mapped jointly without harming agreement (for instance with unweighted κ when a re-mapping does not reach the main diagonal).

2) In the base case where agreement is (strictly) improved by re-mapping a subset of the p_{ij} ratees, it must be optimal to move all the others too, because doing so would increase κ due to linearity in Equation (1).

References

Atlas, R. D. (22 June 2002). Enron's collapse: Credit ratings: Enron spurs debt agencies to consider some changes. *The New York Times*.

Bhanot, K. and Mello, A. S. (2006). Should corporate debt include a rating trigger? *Journal of Financial Economics*, 79(1):69–98.

BIS (2005a). Studies on the validation on internal rating systems. Technical report, Bank for International Settlements. `http://www.bis.org/publ/bcbs_wp14.htm`.

BIS (2005b). Validation of low-default portfolios in the Basel II framework. Technical report, Bank for International Settlements. `http://www.bis.org/publ/bcbs_nl6.pdf`.

BIS (2006). International convergence of capital measurement and capitl standards: A revised framework. Technical report, Bank for International Settlements. `http://www.bis.org/publ/bcbs128.htm`.

Cantor, R. and Packer, F. (1995). The credit rating industry. *Journal of Fixed Income*, 5(3):10–35.

Cantor, R. and Packer, F. (1997). Differences of opinion and selection bias in the credit rating industry. *Journal of Banking & Finance*, 21(10):1395–1417.

Credit Rating Agency Reform Act (2006). 109$^{\text{th}}$ Congress of the United States of America.

ECB, European Central Bank. (2008). General documentation on Eurosystem monetary policy instruments and procedures. *The Implementation of Monetary Policy in the Euro Area*.

Elkhoury, M. (2008). Credit rating agencies and their potential impact on developing countries. Discussion Papers, United Nations Conference on Trade and Development, `http://www.unctad.org/en/docs/osgdp20081_en.pdf`.

Emond, E. J. and Mason, D. W. (2002). A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multicriteria Decision Analysis*, 11(1):17–28.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Frost, C. A. (2007). Credit rating agencies in capital markets: A review of research evidence on selected criticism of the agencies. *Journal of Accounting, Auditing and Finance*, 22:469–492.

Graham, J. R. and Harvey, C. R. (2001). The theory and practice of corporate finance: evidence from the field. *Journal of Financial Economics*, 60(2-3):187–243.

Hornik, K., Jankowitsch, R., Lingo, M., Pichler, S., and Winkler, G. (2007a). Benchmarking credit rating systems. Working paper, Vienna University of Economics and Business Administration, Vienna Graduate School of Finance and Oesterreichische Nationalbank.

Hornik, K., Jankowitsch, R., Lingo, M., Pichler, S., and Winkler, G. (2007b). Validation of credit rating systems using multi-rater information. *Journal of Credit Risk*, 3(4):3–29.

Hornik, K., Jankowitsch, R., Lingo, M., Pichler, S., and Winkler, G. (2010). Determinants of heterogeneity in European credit ratings. *Financial Markets and Portfolio Management*, 24:271–287. 10.1007/s11408-010-0134-x.

Hui, C. H., Wong, T. C., Lo, C. F., and Huang, M. X. (2005). Benchmarking model of default probabilities of listed companies. *Journal of Fixed Income*, 15(2):76–86.

Hunt, J. P. (2009). Credit rating agencies and the "worldwide credit crisis:" the limits of reputation, the insufficiency of reform, and a proposal for improvement. *Columbia Business Law Review*, 2009(1):109.

Jankowitsch, R., Pichler, S., and Schwaiger, W. (2007). Modelling the economic value of credit rating systems. *Journal of Banking and Finance*, 31(1):181–198.

Kemeny, J. G. and Snell, J. L. (1962). *Mathematical Models in the Social Sciences*, chapter Preference Rankings: An Axiomatic Approach. MIT Press.

Kisgen, D. J. (2006). Credit ratings and capital structure. *Journal of Finance*, 61(3):1035–1072.

Kisgen, D. J. (2009). Do firms target credit ratings or leverage levels? *Journal of Financial & Quantitative Analysis*, 44(6):1323–1344.

Kliger, D. and Sarig, O. (2000). The information value of bond ratings. *The Journal of Finance*, 55(6):2879–2902.

Lingo, M. and Winkler, G. (2009). The behavior of erroneous rating systems during changes in the economic environment. Working paper, Vienna University of Economics and Business Administration, Vienna Graduate School of Finance and Oesterreichische Nationalbank.

Lippert, J. (January 2011). Downgraded. *Bloomberg Markets*, p. 87–92.

Lowenstein, R. (27 April 2008). Triple-a failure. *The New York Times*.

Morgan, D. P. (2002). Rating banks: Risk and uncertainty in an opaque industry. *The American Economic Review*, 92(4):874–888.

Pinches, G. E. and Singleton, J. C. (1978). The adjustment of stock prices to bond rating changes. *Journal of Finance*, 33(1):29–44.

Smith, R. C. and Walter, I. (2001). Rating agencies: Is there an agency issue? Working Paper, Stern School of Business, New York University.

Teather, D. (28 January 2003). SEC seeks rating sector clean-up. *The Guardian*.