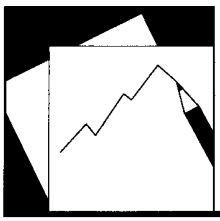
WP/12/122



IMF Working Paper

Appraising Credit Ratings: Does the CAP Fit Better than the ROC?

R. John Irwin and Timothy C. Irwin

INTERNATIONAL MONETARY FUND

 WP/12/122

IMF Working Paper

FAD

Appraising Credit Ratings: Does the CAP Fit Better than the ROC?

Prepared by R. John Irwin and Timothy C. Irwin

Authorized for distribution by Marco Cangiano

May 2012

Abstract

ROC and CAP analysis are alternative methods for evaluating a wide range of diagnostic systems, including assessments of credit risk. ROC analysis is widely used in many fields, but in finance CAP analysis is more common. We compare the two methods, using as an illustration the ability of the OECD's country risk ratings to predict whether a country will have a program with the IMF (an indicator of financial distress). ROC and CAP analyses both have the advantage of generating measures of accuracy that are independent of the choice of diagnostic threshold, such as risk rating. ROC analysis has other beneficial features, including theories for fitting models to data and for setting the optimal threshold, that we show could also be incorporated into CAP analysis. But the natural interpretation of the ROC measure of accuracy and the independence of ROC curves from the probability of default are advantages unavailable to CAP analysis.

JEL Classification Numbers: G24

Keywords: Credit ratings, Receiver Operating Characteristic (ROC), Cumulative Accuracy Profile (CAP).

Authors' E-Mail Addresses: rj.irwin@auckland.ac.nz, tirwin@imf.org

Contents Page

Abstract....1

I. Introduction....3

II. An Illustration: OECD Risk Ratings as Predictors of Borrowing from the IMF....4
A. Cumulative Accuracy Profile (CAP)....5
B. Receiver Operating Characteristic (ROC)....8

III. Four Properties of ROC Analyses not Normally Available to CAP Analyses....9
A. Models....9
B. Theory of Threshold Setting....11
C. Interpretation of Area under the Curve....15
D. Independence from Sample Priors....15

IV. Conclusions....16

Tables
1. Possible Combinations of Predictions and Borrower Behavior....6
2. Frequencies of OECD Rating and Corresponding Rates....7

Figures
1. CAP and ROC Curves for OECD Risk Ratings and Recourse to IMF....6
2. Fitted CAP and ROC Curve....10
3. Indifference Curves and Optimal Thresholds in CAP and ROC Space....14

Appendixes
A. Setting Optimal Thresholds in ROC and CAP Space....17
B. Slope at a Point on a CAP Curve Equals the Likelihood Ratio....20

References....21

I. INTRODUCTION[1]

Judging whether a borrower will repay a loan is a problem central to economic life, and thus assessments of the credit risk posed by borrowers are of great interest. Perhaps the best known assessments are the credit ratings of firms and sovereigns made by Fitch, Moody's, and Standard and Poor's. But there are also credit scores for individuals and credit ratings for firms that are derived from stock prices (see, e.g., Crouhy, Galai, and Mark, 2000). Closely related to credit ratings for sovereigns are ratings of country risk and assessments of the likelihood of fiscal crises (e.g., OECD, 2010; Baldacci, Petrova, Belhocine, Dobrescu, and Mazraani, 2011). Credit ratings not only inform lending decisions, but are also used in rules governing such things as the investments that can be made by pension funds and the collateral that central banks accept. They therefore have an important and controversial influence on financial markets (IMF, 2010).

ROC (Receiver Operating Characteristic) and CAP (Cumulative Accuracy Profile) analyses are two ways of evaluating diagnostic systems. They can be applied to any system that distinguishes between two states of the world, such as a medical test used to detect whether or not a patient has a disease, a meteorological model that forecasts whether or not it will rain tomorrow, and financial analysis that predicts whether or not a government will default on its debt. The key idea underlying ROC and CAP analysis is that diagnosis involves a trade-off between hits and false alarms (that is, between true and false positives) and that this trade-off varies with the stringency of the threshold used to decide whether an alarm is sounded. A good diagnostic system is one that has a high rate of hits for any given rate of false alarms.

Since its introduction in the mid-1950s, the ROC has become the method of choice for evaluating most diagnostic systems, whether in psychology, medicine, meteorology, information retrieval, or materials testing (Tanner and Swets, 1954; Peterson, Birdsall, and Fox, 1954; Swets, 1986). It is not surprising, therefore, that financial analysts have used ROC analysis to assess credit-ratings systems and indicators of financial crisis (e.g., Basel Committee on Banking Supervision, 2005; Engelmann, Hayden, and Tasche, 2003; Sobehart and Keenan, 2001; Van Gool, Verbeke, Sercu, and Baesens, 2011; IMF, 2011). Nevertheless the CAP remains the standard method adopted by financial experts (e.g., Altman and Sabato, 2005; Das, Hanouna, and Sarin, 2009; Flandreau, Gaillard, and Packer, 2010; IMF, 2010; Standard and Poor's, 2010; Moody's, 2009). In this paper, we consider whether the ROC should also become the standard method for appraising credit ratings.

---

[1]We would like to thank Marco Cangiano, Margaret Francis, Michael Hautus, and Laura Jaramillo for valuable comments.

ROC and CAP analyses are similar, and both have the advantage of generating a measure of the accuracy of a diagnostic system that is independent of the choice of diagnostic threshold. Thus both generate a measure of the ability of credit ratings to distinguish between defaulting and nondefaulting borrowers that does not depend on which credit rating is used as the dividing line in any particular application. The reason is that the measures of accuracy take into account all possible thresholds, not just one.

But we show that the ROC has some advantages over the CAP. Because ROC analysis has been widely used for many years, there is a well-known rule for choosing in an ROC setting the diagnostic threshold that maximizes the expected net benefits of the diagnostic decision, given the prior probabilities and the values of hits and false alarms. For the same reason, there is an established body of knowledge about how to fit theoretical ROC models to empirical data. We show, however, how the rule for choosing the optimal threshold and some of the basic theory of model fitting can be translated into the language of the CAP.

Two other advantages of the ROC cannot be transferred so easily to the CAP. First, the principal ROC measure of the accuracy of a diagnostic system has a natural interpretation that the CAP measure of accuracy lacks: if two borrowers are chosen at random, one from the pool of defaulters, the other from the pool of nondefaulters, the probability that the one with the lower credit rating is the defaulter is equivalent to the area under the ROC curve of that ratings system. Second, the shape of the ROC curve, but not the CAP curve, is unaffected by prior probabilities. A rating system's CAP curve therefore changes with the proportion of defaulting borrowers, even when the system's ability to distinguish between defaulters and nondefaulters remains constant. The ROC curve, however, remains the same.

To illustrate the comparison between the ROC and the CAP, we apply these two methods to the Country Risk Classifications made by the Organization for Economic Cooperation and Development (OECD). Our purpose is not to examine OECD ratings, but to present a practical example of the application of these methods in the hope of clarifying the similarities and differences between them.

II. AN ILLUSTRATION: OECD RISK RATINGS AS PREDICTORS OF BORROWING FROM THE IMF

OECD Country Risk Classifications are intended to estimate the likelihood that a country will service its external debt. They are used to set minimum permissible interest rates on loans charged by export-credit agencies and, more specifically, to ensure that those interest rates do not contain an implicit export subsidy. For the purposes of the illustration, we have compared OECD ratings made in early 2002 with a country's recourse to the International Monetary Fund (IMF) during the remainder of the decade, from 2002 to 2010.

It would be possible and, in some respects, more natural to examine how well the ratings of a credit agency predict default. The reason we choose to illustrate the two methods with OECD ratings and IMF lending is not because OECD ratings are intended for that purpose (they are

not), but because this combination provides a straightforward example based on readily available public data. OECD ratings are also available for a larger sample of countries, including many developing countries. And default by governments is much rarer than recourse to the IMF, so a comparison with recourse to the IMF is more informative than comparison with default itself.

We consulted OECD's Country Risk Classifications of the Participants to the Arrangement on Officially Supported Export Credits at http://www.oecd.org/dataoecd/9/12/35483246.pdf. The OECD classifies countries on an eight-point scale from 0 (least risky) to 7 (most risky). We consulted the list compiled between October 27, 2001 and January 25, 2002.

Of 183 countries listed in the IMF's World Economic Outlook Database for October 2010 (http://www.imf.org/external/pubs/ft/weo/2010/02/weodata/weoselgr.aspx), 90 had entered into at least one Fund-supported program during the period between 2002 and 2010 (http://www.imf.org/external/np/pdr/mona). We counted a country as having a program regardless of the type and number of programs accepted during that period.

From the OECD and IMF databases we compiled risk classifications for 161 countries, 82 of which had recourse to an IMF program during the following nine years, and 79 of which did not have recourse to an IMF program.

A. Cumulative Accuracy Profile (CAP)

The left-hand panel of Figure 1 shows the cumulative accuracy profile (CAP) of the OECD ratings in 2002 as predictors of borrowing from the IMF in the following nine years. To construct the CAP curve, we rank countries from riskiest to safest and suppose that each OECD rating is used as a threshold for distinguishing between countries that will subsequently borrow from the IMF and those that will not, and we consider how, as the threshold is varied, the hit rate H co-varies with alarm rate M. The hit rate is the proportion of countries that subsequently borrow from the IMF that are identified as future borrowers, and the alarm rate is the proportion of all countries that are identified as future borrowers. (Table 1 shows the possible outcomes and some of the terminology used in the rest of the paper.[2]) The data points (circles) show the eight OECD risk ratings, from the safest (0) to riskiest (7). Table 2 shows how H and M were computed from the frequency of each rating.

---

[2]There are many variations in terminology. For example, the hit rate and the alarm rate are also called the "true-positive rate" and the "positive rate." In CAP analysis, the ordinate and abscissa of CAP space are sometimes labeled "defaults" and "population" or "cumulative proportion of defaulters" and "cumulative proportion of issuers." In other contexts, the hit rate is called the "sensitivity" and the rate of correct rejections the "specificity."

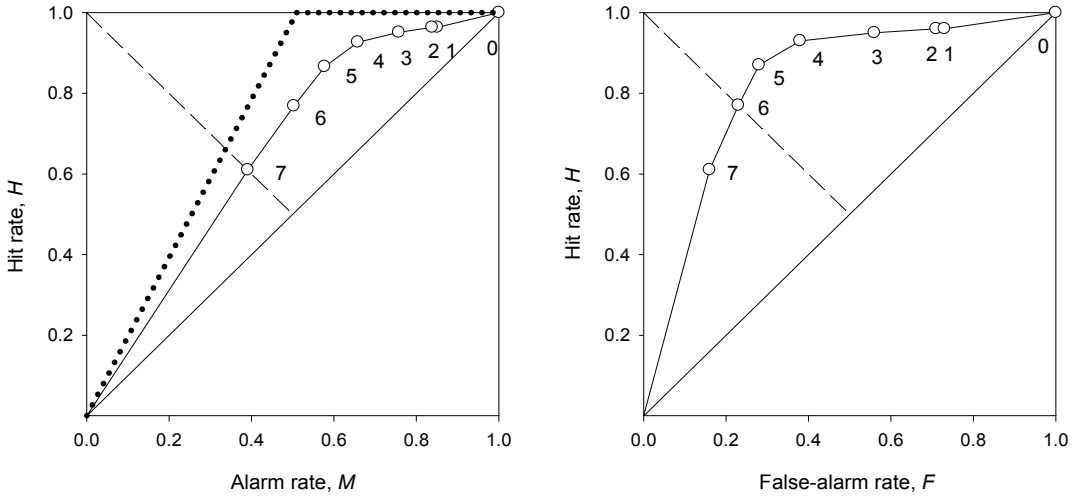
Table 1. Possible Combinations of Predictions and Borrower Behavior

		Country Behavior	
		Borrows from IMF	Does not
		d	*N*
Predicted to borrow (Alarm)	*D*	Hit $\Pr(D\|d;c)=\mathrm{F}_d(c)$	False alarm $\Pr(D\|n;c)=\mathrm{F}_n(c)$
Predicted not to borrow	*N*	Miss $\Pr(N\|d;c)=1-\mathrm{F}_d(c)$	Correct rejection $\Pr(N\|n;c)=1-\mathrm{F}_n(c)$

Note: The symbol *c* denotes a ratings threshold for distinguishing between countries that will subsequently borrow and those that will not, while F_d and F_n denote the cumulative distribution functions of the ratings of borrowers and nonborrowers, respectively.

The hit rate rises with the alarm rate: the greater the proportion of countries that are identified as future borrowers, the greater is the proportion of borrowers that are correctly identified. But, for a given rate of borrowing from the IMF, the steepness of the curve indicates how discriminating the rating system is.

Figure 1. CAP and ROC Curves for OECD Risk Ratings and Recourse to IMF



Note: Left panel: Cumulative Accuracy Profile for OECD Country Risk Classification and subsequent recourse to IMF lending. Each data point (circle), based on a rating from 0 to 7, shows how the hit rate H co-varies with the alarm rate, M. The dotted line shows ideal performance. Right panel: Receiver Operating Characteristic for OECD Country Risk Classification and subsequent recourse to IMF lending. It shows how the hit rate *H* co-varies with the false-alarm rate *F*.

Table 2. Frequencies of Each OECD Rating and their Corresponding Hit Rate (*H*), False-Alarm Rate (*F*), and Alarm Rate (*M*)

IMF Program	OECD Risk Rating							
	0	1	2	3	4	5	6	7
(a) Frequencies								
Yes (Y)	3	0	1	2	5	8	13	50
No (N)	21	2	12	14	8	4	5	13
Sum (T)	24	2	13	16	13	12	18	63
(b) Probabilities cumulated from right to left								
H = cum pr(Y)	1.00	.96	.96	.95	.93	.87	.77	.61
F = cum pr(N)	1.00	.73	.71	.56	.38	.28	.23	.16
M = cum pr(T)	1.00	.85	.84	.76	.66	.58	.50	.39

An index of the performance of a rating system derived from the CAP curve is the accuracy ratio, AR ($-1 \leq AR \leq 1$). It is given by the ratio of two areas: one, Q, is the area bounded by the curve for ideal performance (the dotted line in Figure 1) and the positive diagonal of the unit square. This area indicates the superiority of ideal performance over random performance. The other area, R, is the area bounded by the observed CAP curve and the positive diagonal. This area indicates the superiority of the observed performance over random performance. The ratio of these two areas, R/Q, thus indicates how well the observed performance compares to ideal performance. We show below how this accuracy ratio can also be derived from the ROC curve.

To compute the accuracy ratio for the CAP curve in Figure 1, we first calculate the area S, the proportion of the unit square that lies under the CAP curve. When the data points are joined by straight lines, as in Figure 1, S can be computed by the trapezoidal rule, which gives $S = 0.659$. The area R is then given by $R = S - 0.5 = 0.159$. If the probability of recourse to the IMF is denoted p, the triangular area Q is then given by $Q = \frac{1}{2}(1)(1-p) = \frac{1}{2}(79/161) = 0.245$. Hence the accuracy ratio = $AR = R/Q = 0.65$.[3]

[3]By comparison, Standard and Poor's (2010) reported that, for a ten-year horizon, its foreign-currency ratings of sovereigns had an accuracy ratio of 0.84 and its ratings of private companies had an accuracy ratio of 0.69. These accuracy ratios are higher than that of the OECD ratings in predicting recourse to the IMF, but one needs to acknowledge that the OECD ratings were not intended for that purpose.

The CAP curve and the accuracy ratio are closely related to two concepts commonly used in research on income inequality, the Lorenz curve and the Gini coefficient. Some authors equate them (e.g., Basel Committee, 2005 and Standard and Poor's, 2010). The Lorenz curve shows how much of a population's cumulative income accrues to each cumulative proportion of the population, ordered from poorest to richest, and thus shows how equally income is distributed in the population. The Lorenz curve lies on or below the diagonal, but if the population were instead ordered from richest to poorest it would lie on or above the diagonal. The Gini coefficient, G, is commonly defined as the area between the Lorenz curve and the diagonal, divided by the area under the diagonal. That is, $G = (S - .5)/.5 = 2S - 1$. So, given the above definition of the accuracy ratio, the Gini coefficient and the accuracy ratio are related by $G = AR/(1-p)$.

B. Receiver Operating Characteristic (ROC)

The right-hand panel of Figure 1 shows the ROC curve of OECD ratings as predictors of borrowing from the IMF in the following nine years. The curve was constructed by standard methods for rating ROCs (e.g., Green and Swets, 1966, and see Table 2). It shows how the hit rate H for IMF lending co-varies with its false-alarm rate, F, which is the proportion of nonborrowing countries that are falsely identified as borrowers. Thus, the ROC curve is similar to the CAP curve but whereas the CAP curve relates the hit rate to the rate of all alarms the ROC curve compares it with the rate of false alarms.

The area under the ROC curve in Figure 1 when the points are joined by straight lines is 0.823. Englemann, Hayden, and Tasche (2003) proved that the CAP's accuracy ratio and the area under the ROC curve, A ($0 \leq A \leq 1$), are related by the equation $AR = 2A - 1$. Applying this equation to the OECD data yields $AR = 2 \times 0.823 - 1 = 0.65$ to two decimal places, which agrees with the value calculated for the CAP curve.

Despite the differences between CAP and ROC space, the accuracy ratio of CAP analysis can also be computed directly from the ROC curve, and in essentially the same way that it is calculated from the CAP curve. In particular, it is given by the ratio of two areas: one, Q', is the area bounded by the curve for ideal performance—which in ROC space is a line running from (0, 0) to (0, 1) to (1, 1)—and the positive diagonal of the unit square. This area indicates the superiority of ideal performance over random performance. The other area, R', is the area bounded by the observed ROC curve and the positive diagonal, which indicates the superiority of the observed performance over random performance. As in the case of the CAP space, the ratio of these two areas, R'/Q', thus indicates how well the observed performance compares to ideal performance. Now, it can easily be seen that

$$R'/Q' = (A - .5)/.5 = 2A - 1,$$

which is identical to the accuracy ratio of CAP analysis.

III. Four Properties of ROC Analyses not Normally Available to CAP Analyses

We next discuss four advantageous properties of ROC analysis not available to CAP analysis, as it is traditionally applied. We show how two of these advantages—the existence of models for fitting and interpreting ROC curves and a theory for setting optimal decision thresholds—can be applied to CAP analysis. We then discuss two other advantages that cannot be transferred to CAP analysis—the natural interpretation of the primary measure of accuracy in ROC analysis and the independence of ROC curves from the probability of default (or distress).

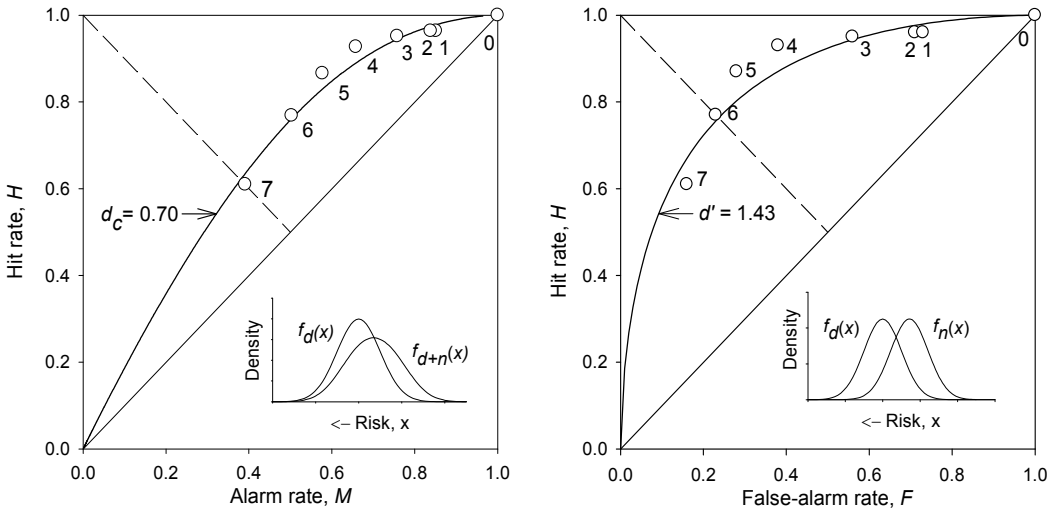
A. Models

A large number of models have been developed for fitting ROC curves to data (see Egan, 1975). For CAP curves there is no such body of knowledge. The right-hand panel of Figure 2 illustrates one such ROC model.

Every detection-theoretic ROC model implies a pair of underlying distributions on a decision variable (or on any monotonic transformation of that decision variable). In this example, one distribution, $f(x|d)$, is conditional on countries' having recourse to IMF lending (d), and one, $f(x|n)$, is conditional on countries' not having recourse to IMF lending (n). We denote these distributions as $f_d(x)$ and $f_n(x)$ respectively. The ROC shows how H and F co-vary with changes in the decision threshold between one rating and the next. When risk decreases with x, the hit rate $H = \mathrm{F}_d(c)$ and the false-alarm rate $F = \mathrm{F}_n(c)$, where F_d and F_n are the distribution functions of $f_d(x)$ and $f_n(x)$ respectively and c is the decision threshold or criterion.

The smooth curve fitted to the data points in the right-hand panel of Figure 2 is based on a standard ROC model, illustrated in the inset, in which the two densities are assumed normal with equal variance. The location parameter of the model is the accuracy index, d', which is the distance between the means of the two densities in units of their common standard deviation. This parameter was estimated to be 1.43 by ordinal regression with IBM SPSS Statistics version 19: it is the location of the mean of the modeled distribution of those countries having recourse to IMF lending relative to the mean of those countries not having such recourse. The area under the normal-model ROC curve is given by $A = \Phi(d'/\sqrt{2})$, where $\Phi(\cdot)$ is the standard normal distribution function (Macmillan and Creelman, 2005). For the ROC in Figure 2, $A = \Phi(1.43/\sqrt{2}) = 0.844$.

Figure 2. Fitted CAP and ROC Curves



Note: Left-hand panel: The smooth curve is the best-fitting normal model to the CAP data from Figure 1, with parameter d_c = 0.70. The inset shows the underlying densities of the fitted model. Right-hand panel: The smooth curve is the best-fitting normal model to the ROC data from Figure 1, with parameter d' = 1.43. The inset shows the underlying densities of the fitted model.

When risk decreases with x, as in the inset of Figure 2, the model fitted to the ROC data can be described by the equation[4] $P(R \geq k \mid J) = \Phi(c_k - Jd')$, where R is an ordinal rating of value k, J is a dummy variable (non-distressed countries = 0 and distressed countries = 1), c_k is the location of the decision threshold for rating k, and d' is the model's accuracy index.

One value of such models is that they can elucidate the nature of the system under study. For example, Irwin and Callaghan (2006) showed how the maximum extreme-value model helped interpret the decision processes of strike pilots who, in a simulated experiment, had to rate whether an emergency warranted ejection. Laming (1986) provided another example. He hypothesized that the shape of a rating ROC curve for detecting brief increments in the energy of light or of sound was determined by the energy distribution of the increments, which is non-central chi-square. He fitted that model ROC to the subjects' ratings of their confidence that they had observed an increment and showed that their decisions were indeed consistent with that hypothesis. We do not attempt an interpretation of the normal model we have fitted to the OECD ratings.

Models of this kind have not to our knowledge been applied to CAP curves. Therefore we next demonstrate how a comparable analysis might be undertaken. Just as every ROC model

[4]cf DeCarlo (2002).

implies a pair of underlying density functions, so too does every CAP model. As above, one probability density function, $f_d(x)$, is conditional on countries' being financially distressed and therefore accepting an IMF program, and another, $f_n(x)$, is conditional on their not being distressed. To model the CAP curve, the weighted sum of these densities is also needed, that is, $f_{d+n}(x) = p \cdot f_d + (1 - p) f_n$ where p is the probability of financial distress. The ordinate of a CAP curve is $F_d(c)$, and the abscissa of a CAP curve is $F_{d+n}(c)$, ordered from riskiest to safest, where F_d and F_{d+n} are the cumulative distribution functions of $f_d(x)$ and $f_{d+n}(x)$ respectively, and c is the decision threshold. A modeled CAP curve then depicts how $F_d(c)$ co-varies with $F_{d+n}(c)$.

The left-hand panel of Figure 2 shows a best-fitting theoretical curve based on two underlying probability densities $f_d(x)$ and $f_{d+n}(x)$ illustrated in the inset. Like the model fitted to the ROC curve, this model has one parameter, which we call d_c: the difference between the means of $f_d(x)$ and $f_{d+n}(x)$. The difference is calculated from the estimated difference between the means of $f_d(x)$ and $f_n(x)$, as described for the ROC curve. When that difference is 1.43, as here, $d_c = 0.703$.

B. Theory of Threshold Setting

Whereas ROC analysis of diagnostic tests stresses the importance of both diagnostic accuracy and threshold-setting, standard CAP analysis yields measures of accuracy only. CAP analysis serves rating agencies well because their primary interest is in accuracy, but lenders and regulators have to make yes-no decisions (e.g., whether to lend or permit lending to a borrower). They therefore need to set thresholds that distinguish safe borrowers from excessively risky ones. One well-known distinction is between "investment grade" ratings (BBB− or higher in the language of Standard and Poor's) and "noninvestment grade" ratings (BB+ or lower). Another is between triple-A and lower ratings.

Analysts sometimes use rules of thumb to choose thresholds. For example, Baldacci, Petrova, Belhocine, Dobrescu, and Mazraani (2011), who developed a new index of fiscal stress for predicting whether a country will experience a financial crisis, considered two rules of thumb. The first is to minimize the total rate of errors (misses and false alarms) or, equivalently to maximize the proportion of correct decisions (hits and correct rejections). Because the false-alarm rate is the complement of the rate of correct rejections, this amounts to maximizing the difference between the hit rate and false-alarm rate, or the vertical distance between the ROC curve and the positive diagonal. This distance is sometimes called the Youden index (see Everitt, 2006) and is closely related to the Pietra index (Lee, 1999). Their second rule of thumb is to maximize the ratio of the hit rate to the false-alarm rate, which they called the "signal-to-noise ratio." This amounts to maximizing the slope of the ROC curve. A third rule of thumb, which is sometimes used to set thresholds for medical diagnosis, is to choose the point on the ROC curve that is closest to perfect performance, namely the point (0, 1). A similar rule would be to select the point nearest (p, 1) in CAP space.

None of these rules of thumb is optimal in general, because they ignore the probability of default and the costs of the two different kinds of error. If the threshold is important, what is needed is not a rule of thumb, but a well-founded theory that incorporates these factors.

The theory for setting decision thresholds is well developed for ROC analysis (see, e.g., Green and Swets, 1966), and Stein (2005) has shown how this theory can be applied to a problem in default prediction. Indeed, principles of decision making formed an important part of the original development of detection theory (Peterson, Birdsall, and Fox, 1954). Among other things, these authors noted that the threshold that maximizes the expected utility of decisions is the one that maximizes the probability-weighted value of the four possible outcomes (Table 1). That is, the optimal threshold, c, is the one that maximizes the following expression (which is derived from Equation A1 in Appendix A).

$$H(c)pV_{Dd} + [1 - H(c)]pV_{Nd} + F(c)(1-p)V_{Dn} + [1 - F(c)](1-p)V_{Nn} \tag{1}$$

where H is the hit rate, F is the false-alarm rate, and p is the probability of default. The values of the four outcomes are denoted V with the subscripts denoting the particular outcome using the notation set out in Table 1. For example, V_{Dd} is the value of the outcome in which default is predicted and happens (or, in our example, borrowing from the IMF is predicted and happens).

Optimal thresholds can be identified by evaluating expression (1). For example, if the probability of borrowing from the IMF is 0.5 and the value of a correct decision about a future borrower ($V_{Dd} - V_{Nd}$) is three times as important as a correct decision about a nonborrower ($V_{Nn} - V_{Dn}$), then the OECD rating threshold that maximizes expected value is 4.

An important aspect of the theory of optimal thresholds is that the slope at any point on the ROC curve is equal to the likelihood ratio of the two underlying distributions that determine that point. Peterson, Birdsall, and Fox (1954) showed that no decision variable could yield better decisions than likelihood ratio. Therefore likelihood ratio is an essential component of optimal decision making. Although this theory is well known, we summarize it in Appendix A in order to compare the results for ROC and CAP curves. There we show (Equation A2) that the threshold that maximizes the expected value of a decision corresponds to the point on the ROC curve for which the likelihood ratio is given by

$$l^R(c^*) = \frac{1-p}{p}\frac{V_{CR-FA}}{V_{H-M}} \tag{2}$$

where $l^R(c^*)$ is the ROC likelihood ratio for the optimal threshold, V_{CR-FA} is the value of a correct rejection relative to the value of a false alarm ($V_{Nn} - V_{Dn}$), and V_{H-M} is the value of a hit relative to the value of a miss ($V_{Dd} - V_{Nd}$). In other words, V_{CR-FA} is the value of making

the right judgment about a non-defaulting or non-distressed borrower and V_{H-M} is the value of making the right judgment about a defaulting or distressed borrower.

As far as we are aware, similar results have not been developed for CAP curves.[5] We first note that, just as the optimum threshold can be found by finding the threshold that maximizes expression (1), which refers to the hit rate, *H*, and false-alarm rate, *F*, of ROC analysis, it can also be found by maximizing the following expression, which refers to the hit rate and alarm rate, *M*, of CAP analysis (see equation A5 in Appendix A):

$$H(c)\,p[(V_{Dd}-V_{Nd})-(V_{Dn}-V_{Nn})]+M(c)(V_{Dn}-V_{Nn})+pV_{Nd}+(1-p)V_{Nn}.$$

We also define the likelihood ratio for the CAP curve since no other decision variable can yield better outcomes. The likelihood ratio for a CAP curve is the likelihood of the density $f_d(x)$ relative to that of $f_{d+n}(x)$ for a given value of x. So the CAP likelihood ratio $l^C(x) = f_d(x)/f_{d+n}(x)$. Now, as shown in Appendix A (Equation A6), the threshold that maximizes the expected value corresponds to the point on the CAP curve for which the likelihood ratio is given by

$$l^C(c^*) = \frac{1}{p}\frac{V_{CR-FA}}{V_{H-M}+V_{CR-FA}}. \tag{3}$$

Just as the slope at a point on an ROC curve equals the likelihood ratio of the threshold that determines that point, so does the slope of a point on a CAP curve equal the likelihood ratio of the threshold that determines that point (see Appendix B). Hence Equation (3) also gives the slope of the CAP curve at the optimal decision threshold.

Although the slopes of both ROC curves and CAP curves equal their determining likelihood ratios, there are important differences between the two curves. For example, whereas the slope of the ROC curve at its optimum can have any positive value (Equation 2), the slope of the CAP curve at its optimum cannot exceed 1/*p* (Equation 3). This maximum is the slope of the line of ideal performance.

In ROC analysis, indifference curves have been used to identify optimal decision thresholds (see, Figure 3, right-hand panel, and Irwin and Irwin, 2011). ROC curves show the feasibility of outcomes, while indifference curves indicate their desirability. Indifference curves can also be used in CAP space, as we illustrate next.

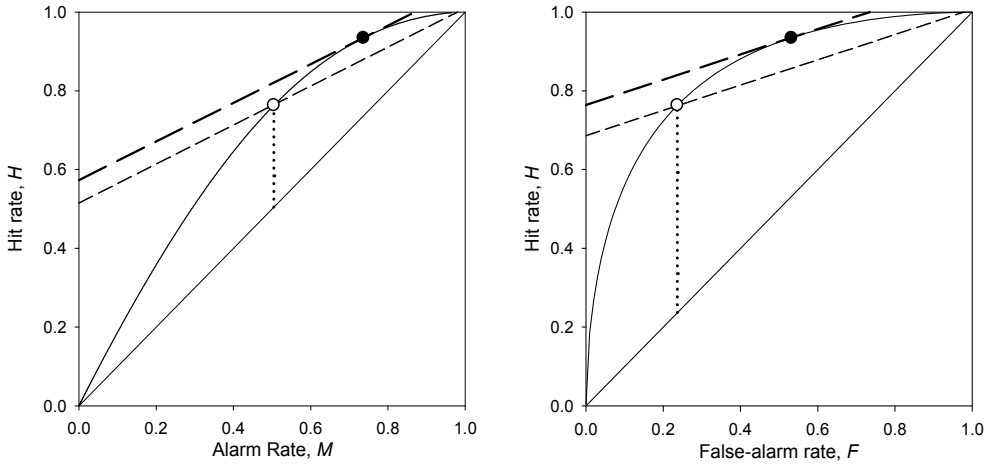
---

[5]Hong (2009) discusses the setting of "optimal" thresholds in both CAP and ROC analyses, but the rules he considers for setting thresholds do not refer to the values of the various outcomes.

Suppose again that a correct decision about defaulters is three times more valuable than a correct decision about non-defaulters and that the probability of default is 0.5. Then the indifference curves in CAP space are straight lines with slope 0.5, as given by Equation (3). The optimal decision threshold for this scenario is then given by the point of intersection between the CAP curve and the highest of indifference curves that intersect the CAP curve. In Figure 3 (left-hand panel) we show the fitted CAP curve and the highest attainable indifference curve (the longer-dashed line), which is tangent to the CAP curve. The optimal threshold is shown by a solid circle in Figure 3. The data points lie above the fitted curve in this region, and the optimal threshold on the fitted curve is nearer to 3 than to 4. The corresponding indifference curve in ROC space is shown by the longer-dashed line in the right-hand panel of Figure 3.

A second decision threshold, selected by the rule of thumb of minimizing the rate of errors, is shown by the open circle in each panel. The dotted vertical lines identify these decision thresholds, which correspond to the maximum vertical distance from the positive diagonal to the CAP or ROC curve. The shorter-dashed lines are indifference curves which pass through these thresholds. As these indifference curves lie below the indifference curve that is tangent to the CAP and ROC curves, the decision thresholds selected by the rule of thumb are, for this scenario, sub-optimal.

Figure 3. Indifference Curves and Optimal Thresholds in CAP and ROC Space



Note: The longer-dashed lines are indifference curves in CAP space (left-hand panel) and ROC space(right-hand panel). The CAP and ROC curves are identical to the fitted curves in Figure 2. Each of these indifference curves is tangent to its corresponding CAP or ROC curve at the optimal decision threshold (see text) shown by the solid circle. The open circles show the sub-optimal points selected by the rule of thumb that minimizes the rate of errors, which locates them at the greatest vertical distance from the positive diagonal, shown by the dotted lines. The shorter-dashed lines are indifference curves passing through these points.

C. Interpretation of Area under the Curve

The area under the ROC curve has been a primary index of accuracy for many years (Green and Swets, 1966). One of its virtues, shared with several other accuracy indices of detection theory, such as the measure d', is that it is a threshold-independent index; that is to say, the area under an ROC curve is independent of the choice of any particular threshold. The threshold-independent property stems from the fact that the area measure is based on all possible thresholds, not on any particular threshold. In this respect it is superior, as a measure of accuracy, to measures such as percentage correct that depend on the threshold chosen. It is also superior to measures, such as the Gini coefficient, that depend on prior probabilities.

The ROC area measure can be interpreted as an unbiased percentage of correct classifications. This interpretation is available by virtue of the area theorem of psychophysics, first proved by Green (1964)—see also Green and Swets (1966) and Egan (1975). The theorem states that the area under a yes–no or rating ROC curve is equal to the proportion of correct decisions of an unbiased observer faced with choosing between the two alternatives analyzed by the ROC. Stated differently, and to continue with our example of IMF lending, if one country is drawn at random from the population of countries that had recourse to the IMF, and another country is drawn at random from the population of countries that did not have recourse to the IMF, then the probability that the country having recourse to the IMF would have a higher risk rating than the country not having such recourse is equal to the area under the corresponding rating ROC. (The theorem is quite general in that it is free of assumptions about the probability distributions that give rise to the ROC.) The area obtained under the ROC for OECD ratings is 0.82, and so by the area theorem these OECD ratings can be said to have an unbiased accuracy of 82 percent—a useful and readily understood interpretation of the result.

The area under the CAP curve, likewise, is threshold independent because it too is based on all the ratings, not a particular rating. And because of the linear relation between A and AR, many of the properties of the ROC area index are shared by the CAP index of accuracy. These include the well-developed statistical properties of the index (e.g., Bamber, 1975).

However, the area theorem does not apply to the CAP accuracy ratio, and so that measure cannot import the readily interpretable meaning as a percentage of correct decisions.

D. Independence from Sample Priors

Measures of accuracy derived from ROC analysis, such as the area under the curve, or the parameter d' of the normal model, do not depend on the sample's prior probabilities because, as Swets, Dawes, and Monohan (2000, p. 26) observed: H and F "are independent of the prior probabilities (by virtue of using the priors in the denominator of their defining ratios). . . . ROC measures do not depend on the proportions of positive and negative instances in any test sample…other existing measures of accuracy vary with the test sample's proportions and are specific to the proportions of the sample from which they are taken."

Several authors have noted the dependence of CAP curves on the composition of the sample. For example, the Basel Committee on Banking Supervision (2005, p. 30) stated: "The shape of the CAP depends on the proportion of solvent and insolvent borrowers in the sample. Hence a visual comparison of CAPs across different portfolios may be misleading." The results of ROC analysis, being independent of the sample composition, therefore enjoy a wider reach than those of the CAP. Despite the dependence of the shape of a CAP curve on the sample priors, its accuracy ratio is independent of them. To see this, note that $AR = 2A - 1$, and because the ROC area A is independent of the sample priors, so therefore is AR.

The dependence of a CAP curve on the composition of the sample is potentially troublesome. For example, the proportion of defaulters or distressed borrowers is much lower in good times than it is in bad times. So CAP curves drawn with data from good times may present a misleading picture of what happens during bad times.

IV. CONCLUSIONS

Both CAP's accuracy ratio and ROC's area under the curve provide threshold-independent indices of accuracy, and both indices are therefore superior to threshold-dependent measures, such as the percentage of correct classifications. Thus, both ROC and CAP analysis provide satisfactory analyses of the accuracy of assessments of OECD risk ratings and other credit ratings. Moreover, the accuracy ratio is linearly related to ROC area under the curve, so the known properties of the ROC measure, such as its statistical properties, are available to CAP results.

However, ROC analyses offer two features not usually reported for CAP analyses. In particular, a well-documented set of models is available for fitting theoretical ROC curves to empirical data, and these in turn may reflect the nature of the underlying rating process. In addition, ROC analysis has an established theory of optimal threshold setting. We have shown how these limitations of CAP analyses might be remedied and how, therefore, the CAP's usefulness in evaluating credit ratings might be improved. We have also shown that the accuracy ratio of CAP analysis is easily computed in ROC space. Two advantages of ROC analysis do not transfer so easily to CAP analysis. The ROC area under the curve has a natural interpretation as the unbiased percentage of correct decisions that does not apply to the accuracy ratio, and the ROC curve, unlike the CAP curve, is independent of the probability of default. We therefore prefer the ROC to the CAP.

Appendix A. Setting Optimal Thresholds in ROC and CAP Space

A credit rater's judgment is either that a borrower will be distressed or default (*D*) or that it will not (*N*). The borrower itself either defaults (*d*) or does not default (*n*). There are thus four possible outcomes (Table 1). Each of the four possible outcomes has a value. Let V_{Dd}, V_{Nd}, V_{Nd}, and V_{Nn} denote those values. The expected value *V* of a using a threshold *c* for distinguishing defaulters from non-defaulters is the probability-weighted sum of these values

$$E[V(c)] = \Pr(D \cap d \,|\, c)V_{Dd} + \Pr(N \cap d \,|\, c)V_{Nd} + \Pr(D \cap n \,|\, c)V_{Dn} + \Pr(N \cap n \,|\, c)V_{Nn}.$$

where $\cap$ denotes intersection, so that $D \cap d$, for example, is the event that the credit rater judges that the borrower will default and the borrower does default. The probabilities are conditional on the threshold *c*, as indicated by notation. By the definition of conditional probability, $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$. Using this fact and letting $p = \Pr(d)$, we have

$$E[V(c)] = \Pr(D \,|\, d;c)\, pV_{Dd} + \Pr(N \,|\, d;c)\, pV_{Nd} + \Pr(D \,|\, n;c)\,(1-p)V_{Dn} + \Pr(N \,|\, n;c)\,(1-p)V_{Nn}.$$

We can write the conditional probabilities in terms of the (cumulative) distribution functions of defaulters and non-defaulters (see Table 1). So we have

$$E[V(c)] = \mathrm{F}_d(c)\, pV_{Dd} + [1-\mathrm{F}_d(c)]pV_{Nd} + \mathrm{F}_n(c)(1-p)V_{Dn} + [1-\mathrm{F}_n(c)](1-p)V_{Nn}. \qquad \text{(A1)}$$

To find the threshold, *c**, that maximizes expected value, we differentiate this equation with respect to the threshold to obtain

$$\frac{d\,E[V(c)]}{d\,c} = \frac{d\,\mathrm{F}_d(c)}{d\,c}\,p(V_{Dd} - V_{Nd}) + \frac{d\,\mathrm{F}_n(c)}{d\,c}(1-p)(V_{Dn} - V_{Nn}).$$

Setting this expression to zero and noting that

$$\frac{d\,\mathrm{F}_d(c)}{d\,c} = f_d(c) \qquad \text{and} \qquad \frac{d\,\mathrm{F}_n(c)}{d\,c} = f_n(c)$$

yields

$$\frac{f_d(c^*)}{f_n(c^*)} = \frac{1-p}{p}\,\frac{V_{Nn} - V_{Dn}}{V_{Dd} - V_{Nd}}.$$

We can simplify the expression by noting that the numerator of the term derived from the values of the possible outcomes is the value of a correct rejection relative to the value of a

false alarm (V_{CR-FA}) and that the denominator of this expression is the value of a hit relative to the value of a miss (V_{H-M}).

We can also note that the ratio

$$\frac{f_d(c)}{f_n(c)} = l^R(c)$$

is the likelihood ratio, where the superscript R refers to ROC, so we can also express the optimality condition as

$$l^R(c^*) = \frac{1-p}{p}\frac{V_{CR-FA}}{V_{H-M}}. \tag{A2}$$

To find the equivalent point on the CAP curve, we note that

$$\mathrm{F}_{d+n}(c) = p\,\mathrm{F}_d(c) + (1-p)\mathrm{F}_n(c)$$

or

$$\mathrm{F}_n(c)(1-p) = \mathrm{F}_{d+n}(c) - \mathrm{F}_d(c)p. \tag{A3}$$

Rearranging (A1) to collect the terms in $\mathrm{F}_n(c)(1-p)$, we get

$$E[V(c)] = \mathrm{F}_d(c)pV_{Dd} + [1-\mathrm{F}_d(c)]pV_{Nd} + \mathrm{F}_n(c)(1-p)(V_{Dn}-V_{Nn}) + (1-p)V_{Nn}. \tag{A4}$$

If we now substitute (A3) into (A4), and rearrange, we get

$$E[V(c)] = \mathrm{F}_d(c)p[(V_{Dd}-V_{Nd})-(V_{Dn}-V_{Nn})] + \mathrm{F}_{d+n}(c)(V_{Dn}-V_{Nn}) + pV_{Nd} + (1-p)V_{Nn}. \tag{A5}$$

Differentiating this with respect to *c*, and setting the resulting equation to zero and solving yields

$$\frac{f_d(c^*)}{f_{d+n}(c^*)} = \frac{1}{p}\frac{V_{Nn}-V_{Dn}}{(V_{Dd}-V_{Nd})+(V_{Nn}-V_{Dn})}$$

or

$$\frac{f_d(c^*)}{f_{d+n}(c^*)} = \frac{1}{p}\frac{V_{CR-FA}}{V_{H-M}+V_{CR-FA}},$$

where the term of the left-hand side of the equation is another likelihood ratio, so we can also write it as

$$l^C(c^*) = \frac{1}{p}\frac{V_{CR-FA}}{V_{H-M} + V_{CR-FA}}. \quad \text{(A6)}$$

To find the equations for indifference curves we follow a similar approach. We set expected utility equal to constants and rearrange the equation so that, for the ROC analysis, the hit rate, F_d, is expressed in terms of the false-alarm rate, F_n, and, for the CAP analysis, the hit rate is expressed in terms of the population F_{dn}.

From Equation (A1), we have the following equation for expected value in ROC terms.

$$E[V(c)] = F_d(c)pV_{Dd} + [1 - F_d(c)]pV_{Nd} + F_n(c)(1-p)V_{Dn} + [1 - F_n(c)](1-p)V_{Nn}.$$

We set this equal to a variable *k*, to get

$$F_d(c)pV_{Dd} + [1 - F_d(c)]pV_{Nd} + F_n(c)(1-p)V_{Dn} + [1 - F_n(c)](1-p)V_{Nn} = k.$$

Collecting the terms in F_d and F_n yields

$$F_d(c)p(V_{Dd} - V_{Nd}) + pV_{Nd} + F_n(c)(1-p)(V_{Dn} - V_{Nn}) + (1-p)V_{Nn} = k.$$

Expressing F_d in terms of F_n then yields the equations for the indifferences curves in ROC space (one for each value of *k*)

$$F_d(c) = \frac{k - pV_{Nd} - (1-p)V_{Nn}}{p(V_{Dd} - V_{Nd})} + \frac{(1-p)(V_{Nn} - V_{Dn})}{p(V_{Dd} - V_{Nd})}F_n(c).$$

We find the equations of the indifference curves in CAP space similarly.

Appendix B. Slope at a Point on a CAP Curve Equals the Likelihood Ratio of the Threshold that Determines that Point

This proof for CAP curves is an adaptation of Green and Swets's (1966) proof for ROC curves. The ordinate of a CAP curve is the cumulative distribution function of the density for defaulters, $f_d(x)$. Its abscissa is the cumulative distribution function of the density for the weighted sum of defaulters and non-defaulters, $f_{d+n}(x)$, where risk x increases from riskiest to safest. Thus for a particular threshold of risk, c, the ordinate is

$$\int_{-\infty}^{c} f_d(x)dx$$

and the abscissa is

$$\int_{-\infty}^{c} f_{d+n}(x)dx\,.$$

To find the slope of a CAP curve at a particular point, we differentiate each of these distribution functions with respect to c:

$$\frac{d}{dc}\int_{-\infty}^{c} f_d(x)dx = f_d(c)\,, \text{ and } \frac{d}{dc}\int_{-\infty}^{c} f_{d+n}(x)dx = f_{d+n}(c)\,.$$

The ratio of the derivative $f_d(c)$ to the derivative $f_{d+n}(c)$ equals the slope, s, of a point on the CAP curve determined by c; and, by definition, it also equals the likelihood ratio at c. That is:

$$s = \frac{f_d(c)}{f_{d+n}(c)} = l^C(c).$$

References

Altman, E. I. and G. Sabato, 2005, "Effects of the New Basel Capital Accord on Bank Capital Requirements for SMEs," *Journal of Financial Services Research,* Vol. 28, pp. 15–42.

Baldacci, E., I. Petrova, N. Belhocine, G. Dobrescu, and S. Mazraani, 2011, "Assessing Fiscal Stress," IMF Working Paper 11/100 (Washington: International Monetary Fund).

Bamber, D., 1975, "The Area Above the Ordinal Dominance Graph and the Area Below the Receiver Operating Characteristic Graph," *Journal of Mathematical Psychology,* Vol. 12: pp. 387–415.

Basel Committee on Banking Supervision, 2005, "Studies on the Validation of Internal Rating Systems," BIS Working Paper No. 14, May, (Basel, Switzerland: Bank for International Settlements).

Crouhy, M., D. Galai, and R. Mark, 2000, "A Comparative Analysis of Current Credit Risk Models," *Journal of Banking and Finance*, Vol. 24, pp. 59–117.

Das, S. R., P. Hanouna, and A. Sarin, 2009, "Accounting-based Versus Market-based Cross-sectional Models of CDS Spreads," *Journal of Banking and Finance,* Vol. 33, pp. 719–730.

DeCarlo, L. T., 2002, "Signal Detection Theory with Finite Mixture Distributions: Theoretical Developments with Applications to Recognition Memory," *Psychological Review,* Vol. 109, pp.710–721.

Egan, J. P, 1975, *Signal Detection Theory and ROC Analysis* (New York: Academic Press).

Englemann, B., E. Hayden, and D. Tasche, 2003, "Testing Rating Accuracy," *Credit Risk*, Vol. 16, January, pp. 82–86.

Everitt, B. S., 2006, *The Cambridge Dictionary of Statistics* (Cambridge: Cambridge University Press, 3rd ed.).

Green, D. M., 1964, "General Prediction Relating Yes-No and Forced-Choice Results," *Journal of the Acoustical Society of America,* Vol. 36A, p. 1042.

————, and J. A. Swets, 1966, *Signal Detection Theory and Psychophysics* (New York: Wiley).

Hong, C. S., 2009, "Optimal Threshold from ROC and CAP Curves," *Communications in Statistics—Simulation and Computation*, Vol. 38, pp. 2060–2072.

International Monetary Fund, 2010, "The Uses and Abuses of Sovereign Credit Ratings," in *Global Financial Stability Report: Sovereigns, Funding, and Systemic Liquidity* (Washington).

————, 2011, "Toward Operationalizing Macroprudential Policies: When to Act?" in: *Global Financial Stability Report: Grappling with Crisis Legacies* (Washington).

Irwin, R. J., and K. S. N. Callaghan, 2006, "Ejection Decisions by Strike Pilots: An Extreme Value Interpretation," *Aviation, Space, and Environmental Medicine,* Vol. 77, pp. 62–64.

————, and T. C. Irwin, 2011, "A Principled Approach to Setting Optimal Diagnostic Thresholds: Where Receiver Operating Characteristic and Indifference Curves Meet," *European Journal of Internal Medicine,* Vol. 22, pp.230–234.

Keenan, S., and J. Sobehart, 2000, "A Credit Risk Catwalk," *Risk*, July, pp. 84–88.

Laming, D., 1986, *Sensory Analysis* (New York: Academic Press).

Lee, Wen-Chung, 1999, "Probabilistic Analysis of Global Performances of Diagnostic Tests: Interpreting the Lorenz Curve-Based Summary Measures," *Statistics in Medicine,* Vol.18, pp. 455–471.

Macmillan, N. A. and C. D. Creelman, 2005, *Detection Theory: A User's Guide* (Mawha, N.J.: Lawrence Erlbaum Associates, 2nd ed.).

Moody's Corporation, 2009, "Sovereign Default and Recovery Rates, 1983–2008," Available via the Internet: http://v2.moodys.com/cust/content/content.ashx?source=StaticContent/Free%20pages/Credit%20Policy%20Research/documents/current/2007400000587968.pdf

Organization for Economic Cooperation and Development, 2010, "Country Risk Classification of the Participants to the Arrangement on Officially Supported Export Credits, 1999–2010." Available via the Internet: http://www.oecd.org/dataoecd/9/12/35483246.pdf

Peterson W. W., T. G. Birdsall, and W. C. Fox, 1954, "The Theory of Signal Detectability," *Transactions of the IRE Professional Group on Information Theory*, Vol 4, No. 4, pp. 171–212.

Sobehart, J., and S. Keenan, 2001, Measuring Default Accurately. Credit Risk Special Report, *Risk*, pp. S31–S33.

Standard and Poor's Corporation, 2010, "Sovereign Defaults and Rating Transition Data, 2009 Update." Available via the Internet: http://www.standardandpoors.com/ratings/articles/en/us/?assetID=1245228008798.

Stein, R. M., 2005, "The Relation Between Default Prediction and Lending Profits: Integrating ROC Analysis and Loan Pricing," *Journal of Banking & Finance,* Vol. 29, pp. 1213-1236.

Swets, J. A., 1986, "Form of Empirical ROCs in Discrimination and Diagnostic Tasks: Implications for Theory and Measurement of Performance," *Psychological Bulletin,* Vol. 99, pp.181–198.

————, R. M. Dawes, and J. Monohan, 2000, "Psychological Science Can Improve Diagnostic Decisions," *Psychological Science in the Public Interest,* Vol. 1, pp. 1–26.

Tasche, Dirk, 2006, "Validation of Internal Ratings Systems and PD Estimates," Available via Internet: http://arxiv.org/abs/physics/0606071.

Tanner, W.P. Jr., and J. A. Swets, 1954, "A Decision-Making Theory of Visual Perception," *Psychological Review,* Vol. 61, pp. 401–409.

Van Gool, J., W. Verbeke, P. Sercu, and B. Baesens, 2011, "Credit Scoring for Microfinance: Is it Worth it?" *International Journal of Finance and Economics,* Vol. 17, No. 2, pp. 103–123