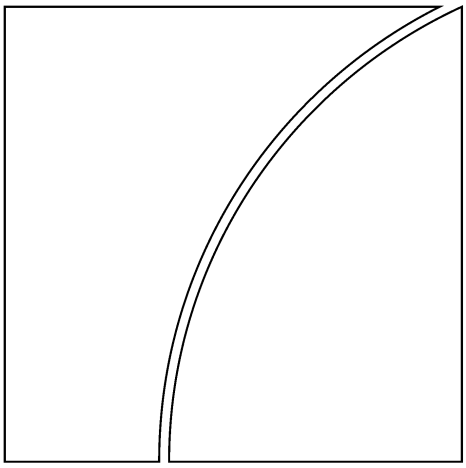
BANK FOR INTERNATIONAL SETTLEMENTS



BIS Working Papers

No 133

How good is the BankScope database? A cross-validation exercise with correction factors for market concentration measures

by Kaushik Bhattacharya

Monetary and Economic Department

September 2003

BIS Working Papers are written by members of the Monetary and Economic Department of the Bank for International Settlements, and from time to time by other economists, and are published by the Bank. The views expressed in them are those of their authors and not necessarily the views of the BIS.

Copies of publications are available from:

Bank for International Settlements
Press & Communications
CH-4002 Basel, Switzerland

E-mail: publications@bis.org

Fax: +41 61 280 9100 and +41 61 280 8100

This publication is available on the BIS website (www.bis.org).

ISSN 1020-0959 (print)

ISSN 1682-7678 (online)

Abstract

The paper examines the quality of the BankScope database, by comparing results based on it to those obtained from the population-level data for India disseminated by the Reserve Bank of India. Despite good coverage and minor reporting errors in the individual reported units, strong evidence of selectivity bias in BankScope data for India is found. A major source of the selectivity bias for India is the almost total omission of Regional Rural Banks and Foreign Banks. It is shown that this selectivity bias affects estimates of all summary statistical measures and could lead users of the data to conclude that the Indian banking market is unimodal when, in reality, it is segmented and has a bimodal pattern. Kolmogorov-Smironov tests reveal that neither the distribution of the log of total assets nor that of market shares based on the BankScope data could be treated same as the corresponding population distributions for India. Despite these limitations, the paper shows that a few popularly used market concentration measures could be estimated from BankScope data accurately, provided the coverage ratio with respect to the size variable is known from alternative sources and is adequate. Coverage of about 90% with respect to the size variable is found to be sufficient for approximating population HHI. For k-bank concentration measures, accurate estimates could be obtained if, in addition, the top k banks in the population are also available in the sample. In contrast, for entropy measures, results indicate that adequate coverage with respect to both the size variable and the number of financial entities would be required.

Keywords: Kolmogorov-Smironov test, market concentration measures.

Journal of Economic Literature Classification: C8, D4.

Table of contents

1. Introduction[1]

The BankScope database is a unique collection of micro-level banking information for different countries. It is used by many leading financial institutions, including central banks, for cross-country studies and policymaking (Demirgüç-Kunt and Detragiache (1998), De Bandt and Davis (1999), Corvoisier and Gropp (2001)). The BankScope database, however, is a sample and does not cover the entire population of banks in a country. It is, therefore, essential to examine how good and representative these samples are for different countries. Unfortunately, in the absence of a comparable database, such validation exercises are rare and are typically focused on coverage with respect to a size variable. For example, Cunningham (2001), in a recent study spanning 19 emerging market economies, observed that, in 15 of these markets, BankScope data covered more than 90% of the total banking sector assets.[2]

Comparison of the coverage with respect to total assets is, however, one small aspect of overall data validation.[3] It is well known that banking market structures in many economies are heterogeneous and segmented. To assess the true quality of a sample, it is, therefore, necessary to examine whether the sample distributions with respect to specific size variables are representative. Thus, it is important to juxtapose the distribution of total assets of individual banks in the sample with the overall population distribution for different countries and, through this, to test whether there exists any selectivity bias in BankScope data. Further, in many cases, economists and policymakers routinely require and estimate some summary measures (eg, market concentration measures) from these distributions for policy purposes as well as for cross-country comparisons. Hence, cross-validation of this aspect is also essential. While a few earlier users of BankScope data like De Brandt and Davis (1999) and Corvoisier and Gropp (2001) raised some suspicions about the possibility of selectivity bias in the data, no rigorous analysis has yet been done on the subject. Any validation of the data set would, therefore, put the empirical inferences drawn from these studies on firmer ground and, alternatively, any exposure of weakness in the data, as suspect.

The purpose of this paper is twofold. First, it attempts a detailed validation exercise on BankScope data for India. Fortunately, the recent public release of detailed micro-level banking data in India has made this study possible.[4] Though the validation exercise is restricted to India due to constraints on availability, it is expected that the exercise attempted here would set a benchmark for similar exercises for other countries. The study compares the existing sample distribution of bank assets based on BankScope data to that obtained from the population data disseminated by the Reserve Bank of India (RBI). It also examines to what extent popularly used statistical measures are affected in the BankScope database. The paper, in this context, finds strong evidence of selectivity bias in BankScope data for India. The second purpose of the paper is to demonstrate that, with some additional information, the selectivity bias in a few popularly used market concentration measures from BankScope data could still be corrected. The correction factors derived are general and are based on

---

[1] The research reported in this paper was carried out when the author was a Visiting Research Fellow at the Bank for International Settlements (BIS) from the Reserve Bank of India (RBI). The author would like to thank Konstantinos Tsatsaronis and Madhusudan Mohanty for discussions and comments on an earlier draft and Angelika Donaubauer, Janet Plancherel and Tom Minic for help of different kinds. The incorporation of some of the suggestions of participants in seminars at both the BIS and the RBI have also substantially improved the earlier exposition. The views expressed in the paper are purely personal and not those of the BIS or the RBI. The responsibility for any errors that remain lies solely with the author.

[2] The countries considered by Cunningham (2001) were Argentina, Brazil, Chile, Colombia, Venezuela, Mexico, China, India, Indonesia, Korea, Malaysia, the Philippines, Taiwan, Thailand, the Czech Republic, Hungary, Poland, Russia and Turkey. Countries for which coverage was found to be less than 90% were China, Indonesia, the Czech Republic and Russia. The reference year was 1999.

[3] Cunningham (2001) acknowledged the possibility that such an exercise was "only indicative". However, rather than selectivity bias, his main concern was that there might be differences in reporting treatment between local sources and the bank's account within the database.

[4] The data are publicly disseminated by the RBI on its website (http://www.rbi.org.in). The first set, entitled "Annual Accounts Data of Scheduled Commercial Banks (1989-90 to 2000-01)", disseminates panel data and the second, entitled "Statistical Tables Relating to Banks in India", provides data pertaining to the most recent financial year.

publicly available macro-level information. The paper demonstrates that if the coverage ratio is reasonably high for a country, use of the correction factors could yield better estimates of market concentration in that country based on BankScope data.

The paper is organised as follows. Section 2 juxtaposes the sample and the population distributions and examines the implications for a few popularly used statistical measures. Section 3 is specifically devoted to the implications for concentration measures. Section 4 discusses in brief under what conditions the selectivity bias could be reduced or eradicated for market concentration measures. Finally, Section 5 summarises the findings with a few concluding observations.

2. Comparison of statistical features

Compared to many other emerging market economies, India has an extensive banking network. The scheduled banking structure in India consists of banks that are listed in the Second Schedule of the Reserve Bank of India Act, 1934. These scheduled banks are divided in two groups: (i) Scheduled Commercial Banks (SCBs) and (ii) Scheduled Cooperative Banks. This study is restricted to the SCBs, which account for more than 90% of banking business in India.[5] For analytical purposes, the SCBs can be further classified into four major groups: (i) Public Sector Banks, (ii) Indian Private Sector Banks, (iii) Regional Rural Banks (RRBs) and (iv) Foreign Banks (FBs). Among the Public Sector Banks, official reports generally indicate results separately for: (i) State Bank of India (SBI) and Its Associates and (ii) Other Nationalised Banks, due to the large size of the SBI. This paper also maintains the distinction, and henceforth, works with five bank groups.

The organisational structure of the BankScope CD (Update 150.1, January 2003; henceforth referred to as the BankScope CD), however, appears to be slightly different and more tuned towards international comparisons. That is why, from this database, banks could be filtered according to their country of operation as well as according to specialisation. Thus data on a specific country cover both locally owned banks and foreign-owned subsidiaries. The coverage of foreign entities is important because existing evidence seems to suggest that foreign participation in emerging market economies appears to be increasing (Mathieson and Roldos (2001)). The specialisation categories in the BankScope database are: (i) Commercial Banks, (ii) Savings Banks, (iii) Cooperative Banks, (iv) Real Estate and Mortgage Banks, (v) Medium-Term and Long-Term Credit Banks, (vi) Investment Banks and Securities Houses, (vii) Islamic Banks, (viii) Non-Banking Credit Institutions, (ix) Specialised Governmental Credit Institutions, (x) Bank Holdings and Holding Companies, (xi) Central Banks, and (xii) Multilateral Government Banks. Two aspects are observed from this classification. First, the groups in the above classification scheme are not always mutually exclusive. Second, besides banks, the database also includes other financial entities. Interestingly, any discussion on the detailed aspects of survey design or inclusion policy was conspicuously missing from the BankScope CD.

In this study, the empirical comparison has been restricted to the year 2001. It may be noted that the BankScope CD contains detailed annual time series data from 1991 onwards. The RBI CD entitled "Annual Accounts Data of Scheduled Commercial Banks (1989-90 to 2000-01)" also covers the same period. However, the coverage of the BankScope CD for India during the early 1990s appears to be weak. So far as the latter half of the 1990s is concerned, empirical evidence appears to suggest that, despite a spate of mergers, the market structure of banking in India did not undergo significant changes (Bhattacharya and Das (2003)). The choice of a single year is, therefore, perceived as adequate.

The filtering pertaining to India isolated 103 financial entities in the BankScope database. Among these entities, those for which exact matches were obtained in the population data were identified. The number of such entities was 62, for 58 of which data on total assets were available during the financial year 2000\01 for final analysis.[6] Further examination of the data yields interesting

---

[5] The comparison is restricted to the SCBs primarily due to lack of availability of detailed micro-level balance sheet and profit and loss account data on the Scheduled Cooperative Banks.

[6] The financial entities whose names matched in the list of the SCBs of the RBI but showed no data for 2001 were Benares State Bank, Crédit Lyonnais (Indian branches), Ganesh Bank of Kurundwad and Standard Chartered Bank (Indian branches).

results. Balance sheets and profit and loss accounts data in India pertain to a financial year, which lasts from the beginning of April in a given year to the end of March in the next calendar year. However, annual data isolation through BankScope on a calendar year basis reports the "nearest" result available to the calendar year. Thus, if one wants to match the figures pertaining to "end-March 2001" from BankScope data for India, one needs to filter results for the year 2000 in the database, and not 2001.

In Table 1, we compare the results obtained from the BankScope database to the population across bank groups. The results reveal interesting features. Though BankScope data cover less than one-fifth of the total number of SCBs in India, in terms of total assets, their coverage is 88.99%. However, a detailed breakdown of coverage in terms of bank groups reveals a high degree of variation in terms of coverage. Among the five major groups in India, the nationalised banks – comprising (i) SBI and Its Associates and (ii) Other Nationalised Banks – are fully represented in the BankScope database. Although the aggregate figure for total assets matches almost totally to that for the first group, a comparison for group (ii) reveals some discrepancies. In absolute terms, the BankScope database underreports the total assets of this group by about INR 78,394 million. A detailed examination of the source of discrepancy reveals that there is almost total agreement for 12 banks within this group. The source of the deviation is primarily due to 7 banks. Table 2 presents the figures for the total assets of these 7 banks as reported in both the databases. It is found that the total deviation due to these 7 banks is almost exactly equal to the figure of total deviation for the group as a whole. Among these, the Indian Bank alone accounts for about half of the total deviation. The deviations for UCO Bank and United Bank of India are also high. The deviations for Dena Bank and Central Bank of India are moderately positive, while those for Canara Bank and Punjab National Bank are comparatively low but negative. It may be noted that some of these banks are troubled banks. Hence, it is likely that some redefinitions of total asset sizes of these banks were done in the BankScope database to make the figures consistent with international standards.

Table 1 also reveals that the Indian Private Banks are also very well represented in the BankScope database. BankScope data cover 29 out of 31 such banks reported by the RBI. The two banks ignored in the BankScope database are Benares State Bank and Ganesh Bank of Kurundwad, the total asset sizes of the two being INR 11,346 million and INR 1,754 million respectively. It may be noted that together they account for INR 13,100 million of the total reported gap of INR 13,656 million in terms of coverage in this group. The deviations in the reported figures within this group from the original figures were minor. For the majority of banks, the figures agreed totally; when they did not, the deviations were generally within INR 100 million. As the two banks ignored in the BankScope database cover only a minor percentage of total assets within the group and a minuscule fraction of total bank assets in India, the omission is not serious.

Table 1
Comparison of bank assets across bank groups

Bank groups	Number of banks		Bank assets	
	BankScope	Population	BankScope	Population
(i) SBI and Its Associates	8	8	4,028,771.5	4,028,770
(ii) Other Nationalised Banks	19	19	16,190,525.9	6,268,920
(iii) Regional Rural Banks	0	196	0.0	495,960
(iv) Indian Private Banks	29	31	1,620,144.2	1,633,800
(v) Foreign Banks	2	42	126,162.9	1,018,240
All Scheduled Commercial Banks	**58**	**296**	**11,965,604.5**	**13,445,770**

Note: Amounts are in millions of Indian rupees (INR). Due to rounding, individual values may not agree with the totals.

However, the major limitation of coverage in the BankScope database appears in covering groups (iii) and (v), ie the RRBs and the FBs operating in India. The first provides an important micro-aspect of Indian banking, while the role of the second is becoming increasingly important in view of the globalisation of financial markets. It may be noted that FBs could propagate financial crisis through contagion. BankScope data do not cover group (iii) at all, while for group (v), their coverage appears to

be grossly inadequate. Consistent with the general pattern observed by Mathieson and Roldos (2001), it is well known that in India the overall contribution of FBs in terms of total assets is increasing. The total assets of FBs experienced a sharp increase from INR 828,500 million in end-March 2000 to INR 1,018,240 million in end-March 2001, an increase of about 22.9%. Thus, it is perceived that the overall coverage of BankScope data for India deteriorated to 88.99% from the figure of more than 90% observed by Cunningham (2001) due to lack of adequate coverage of FBs.

Table 2
Major deviations between reported figures of total bank assets of a few Nationalised Banks in the BankScope database and the RBI

Bank (1)	**Report total asset**		**Deviation (4) = (3) – (2)**
	BankScope (2)	**RBI (3)**	
Canara Bank	665,206	664,390	–816
Central Bank of India	465,786	472,603	6,817
Dena Bank	176,713	179,086	2,373
Indian Bank	227,577	266,405	38,828
Punjab National Bank	635,192	635,051	–141
UCO Bank	255,570	273,312	17,742
United Bank of India	201,236	214,828	13,592
Total			**78395**

Note: Amounts are in millions of INR.

An important question that needs to be answered in this context is: how serious is the omission of RRBs and FBs? If the FBs in India, as a group, display similar patterns of asset distribution as the domestic banks, clearly the omission is not serious. Hence a detailed comparison of the major statistical features in the two databases is essential.

In all subsequent analysis, two sets of exercises are carried out: including and excluding the RRBs. The distinction is maintained because sometimes policy analyses by the official agencies in India exclude RRBs from the domain. The distributional structures of the log of assets of the three data sets, viz, BankScope data and the two sets of "population data", respectively including and excluding the RRBs, are examined. The summary statistical measures pertaining to the log of asset values for these three data sets are presented in Table 3.

Table 3
Comparison of statistical properties of the log of asset values

Statistical properties	**BankScope**	**Population (with RRBs)**	**Population (without RRBs)**
Mean	11.4056	8.4962	10.2533
Median	11.5572	7.9771	10.5577
Standard deviation	1.3250	1.7996	1.9880
Coefficient of variation (%)	11.6171	21.1812	19.3889
Skewness	-0.2162	1.1673	-0.2443
Kurtosis	0.2392	0.6430	-0.7675
Observations	58	296	100

As the assets of individual banks have been converted to log values, the difference reported in Table 3 could be substantial. As expected, BankScope reports a comparatively high mean and median and less standard deviation. Statistical tests reject the equality of the mean and the standard deviation based on the BankScope data to the respective population values, irrespective of inclusion or exclusion of the RRBs in the population. In fact, one-way tests reveal strong evidence that the mean of the log of assets based on the BankScope data could be higher than the population values, indicating the presence of selectivity bias. The bias is high if the RRBs are considered within the domain of analysis. More importantly, the higher moments seem to diverge considerably. In the case of skewness and kurtosis, even the signs in the BankScope sample are sometimes not consistent with the population.

Figure 1
Frequency distribution of the log of assets for all banks (including RRBs)

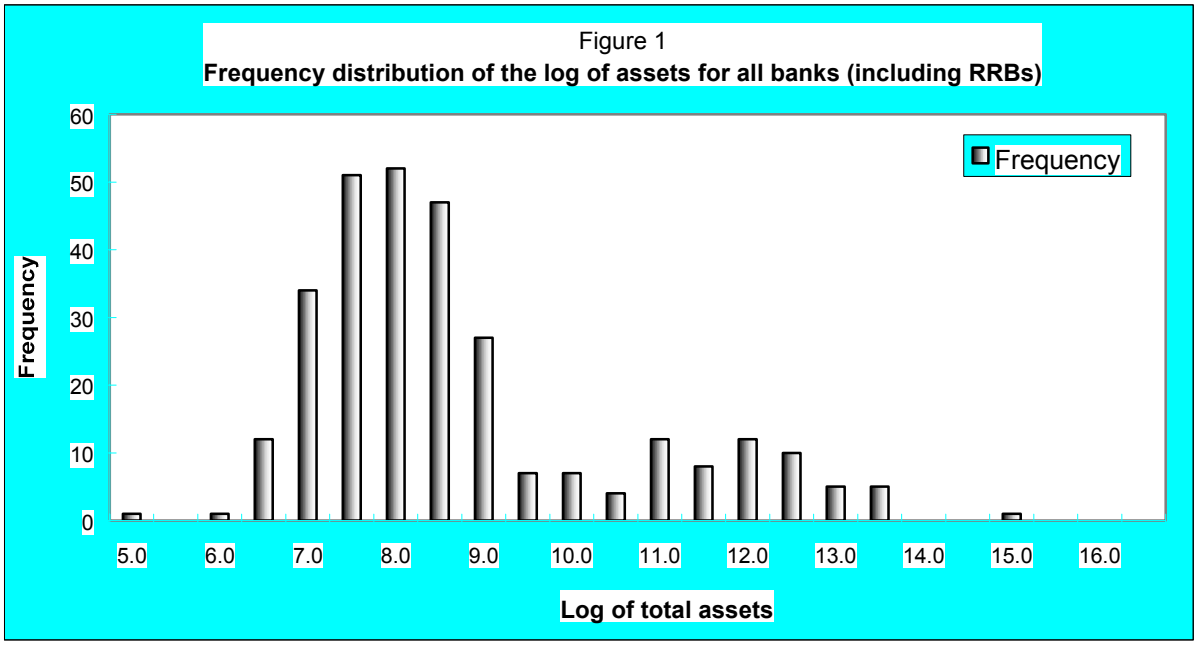


Figure 2
Frequency distribution of the log of assets for all banks (excluding RRBs)

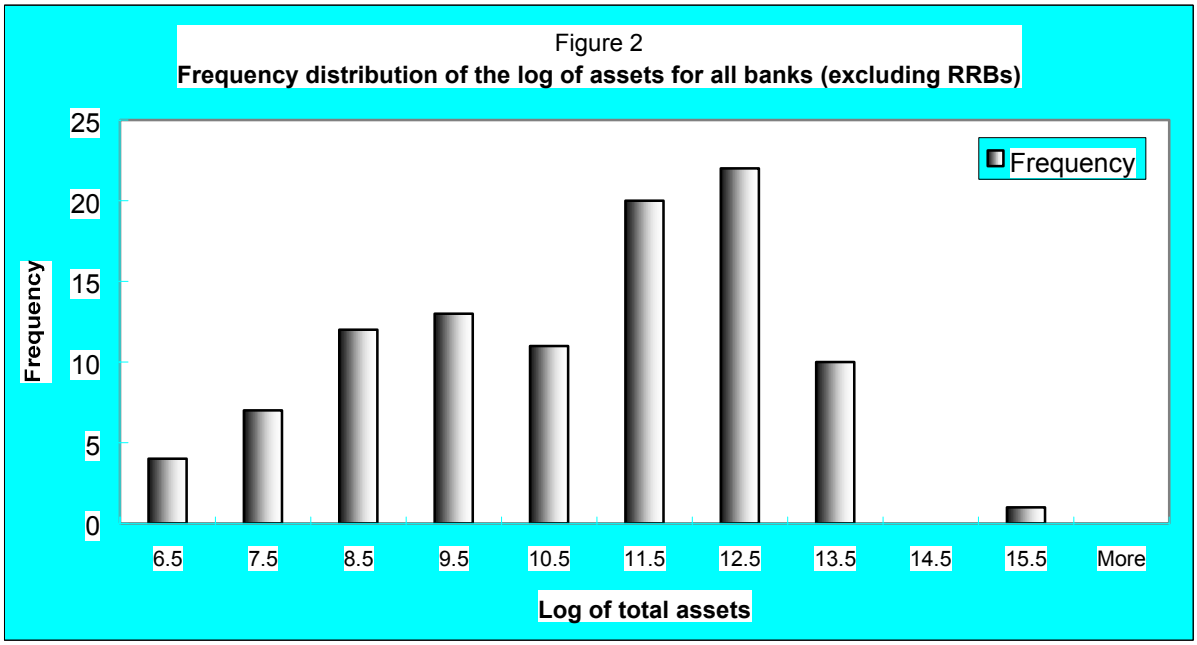
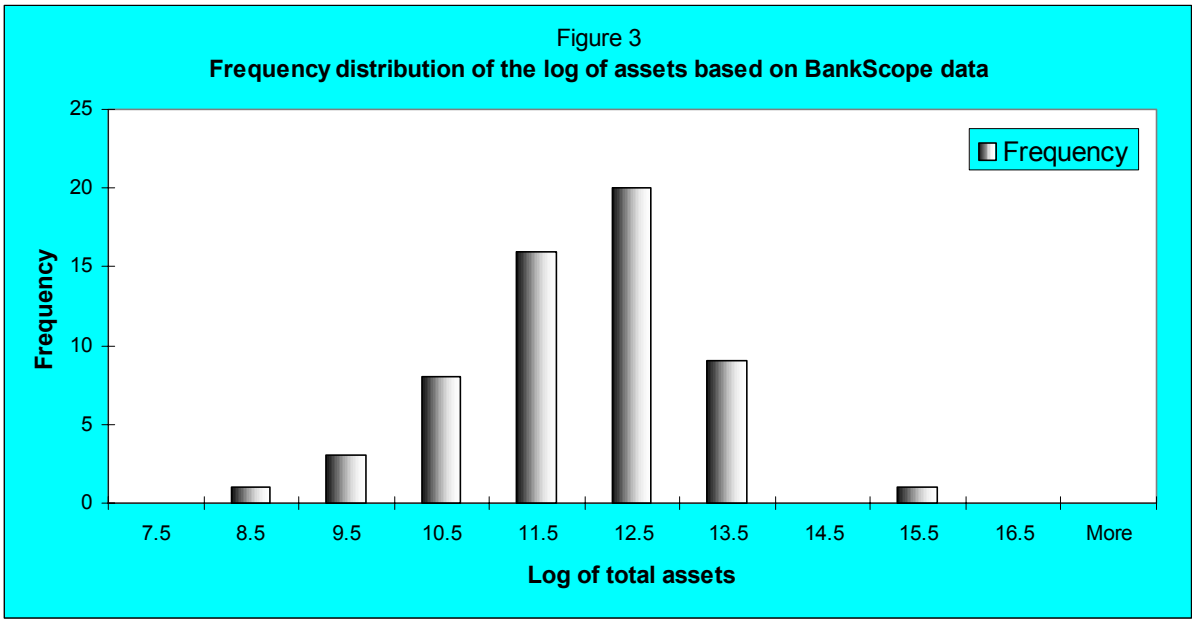


Figure 3

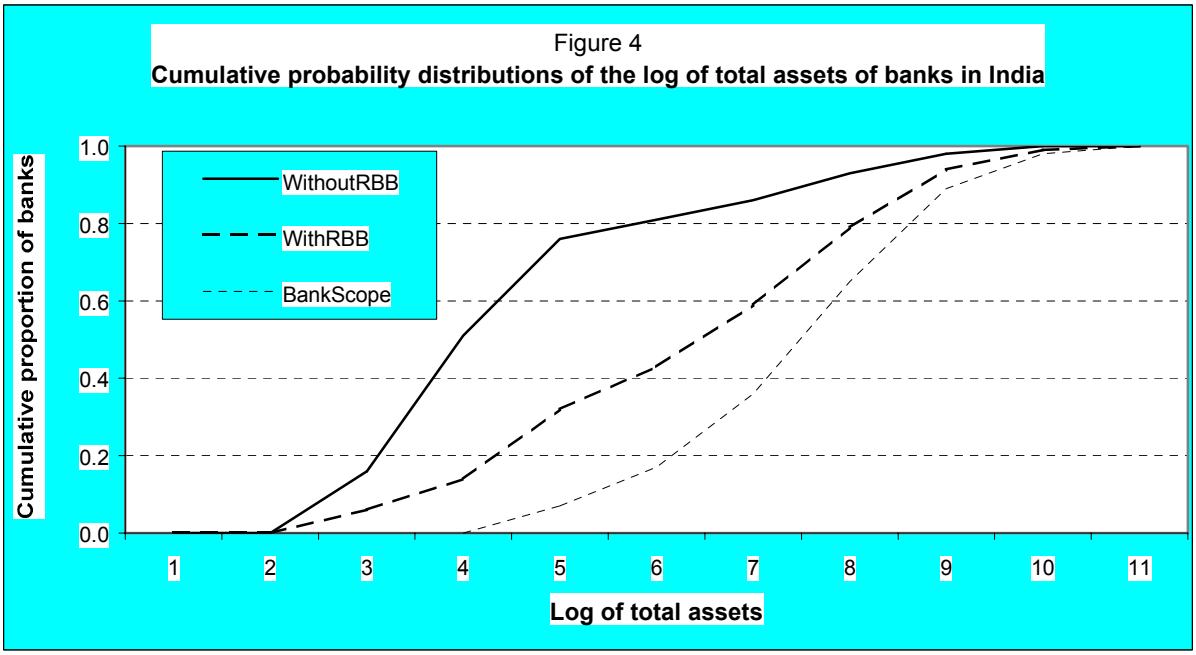
Frequency distribution of the log of assets based on BankScope data



The full frequency distributions pertaining to the three data sets are presented in Figures 1–3. Note that both Figures 1 and 2 clearly indicate the presence of a mixture distribution. Figure 1 appears to be more flatly distributed, with a good number of banks at the tails. The population figures indicate a bimodal shape. The bimodal shape is more evident if RRBs are excluded from the domain of population, as in Figure 2. The figures thus indicate that, in the Indian context, FBs form an important segment and have special statistical features. BankScope data, by not representing this group of banks sufficiently in the database, would lead the user to think that the Indian banking structure is unimodal.

Figure 4

Cumulative probability distributions of the log of total assets of banks in India



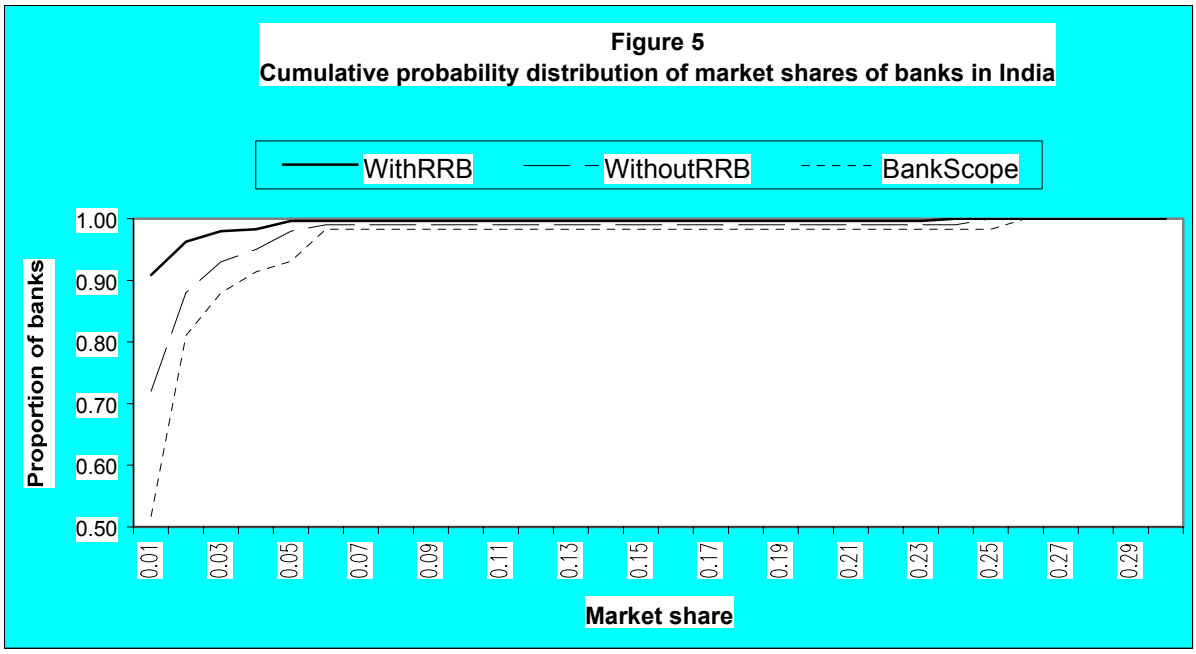
In Figure 4, we plot the empirical CDFs corresponding to the log of total bank assets for the three data sets. Figure 4 reveals wide divergence between the empirical CDFs as obtained from BankScope and

the population, whatever the definition of population. The one-sample Kolmogorov-Smironov test statistics corresponding to two tests of equality are 0.69 and 0.26 when RRBs are included and excluded respectively. It is well known that for a sample size greater than 35, the critical value at the 0.05 level for the one-sample K-S test is $1.36/\sqrt{n}$. Here n equals 58, implying that the critical value turns out to be 0.1786. As both the values of the test statistics are higher than the critical values, the tests reveal strong evidence that whatever the definition of the population, BankScope data do not present the distributional structure of bank assets in India correctly.

3. Comparison of concentration measures

In this section, the performance of BankScope data vis-à-vis the RBI data is examined for market concentration measures. Market concentration measures are often used by policymakers to assess the possible impacts of mergers or acquisitions.[7] Given the recent worldwide emphasis on adopting an appropriate competition policy framework, especially in the context of emerging market economies, the importance of having correct values pertaining to these measures has increased (Neumann (2001), Singh (2002)). What increases the relevance of the study in the present context is that sometimes the BankScope data have been used to measure banking market concentration across countries (Demirgüç-Kunt and Detragiache (1998), Corvoisier and Gropp (2001)).

Figure 5
Cumulative probability distribution of market shares of banks in India



In the case of validating popularly used market concentration measures, the comparison of probability distribution is not based on the log of asset size, but on the distribution of market shares. Figure 5 plots the cumulative probability distribution of market shares of banks based on the three available data sets. The lines corresponding to WithRRB and WithoutRRB depict the two population CDFs including RRBs and excluding RRBs respectively. The BankScope line traces the empirical CDF as obtained from the BankScope data. It is observed that the distributions are highly dissimilar for lower values of market share. The values of Kolmogorov-Smironov D statistics are 0.3916 and 0.2028 in the

[7] For example, the antitrust policy framework in the United States is guided by the existing level of and the possible changes in a few common market concentration measures.

two cases, when the "population" includes and excludes RRBs respectively. The fact that these values are greater than the critical value (0.1786 here) implies that the tests find strong evidence that the distribution of market share as obtained from the BankScope data cannot be treated as identical to any of the population distributions.

Despite the difference in distributional structure, it would be interesting to examine how well BankScope data approximate the market concentration measures. It is well known that different indices of concentration put different weights on different parts of the distribution of market shares across firms and may give contradictory evidence. Let there be *n* firms in an industry with market shares $s_1, s_2, \ldots, s_n$. A simple but general linear form of an index of industrial concentration (IIC) is:

$$\text{(1)} \qquad \text{IIC} = \sum_{i=1}^{n} w_i s_i$$

where w_i ($i=1,2, \ldots, n$) are weights that may or may not sum to unity.

Following the taxonomy of Marfels (1971), there could be four broad classes of weights: (i) unity to top k banks and zero to the rest (example: k-bank concentration ratios, denoted by CR_k), (ii) individual ranks of banks (example: Hall-Tildeman Index, denoted by HTI), (iii) banks' own market shares or their power (example: Herfindahl-Hirschman Index, denoted by HHI, which uses banks' own market shares as weights), and (iv) the negative of the logarithm of market shares (example: entropy measure). The different schemes reflect different assessments regarding the relative impact of larger and smaller firms. Depending upon the scheme, the various concentration measures could show strongly divergent values for the same market. However, though the individual measures may vary, they often lead to similar orderings over space or time.[8] Table 4 presents the taxonomy of Marfels (1971) in some detail, displaying the exact functional form of a few popularly used market concentration measures, along with their properties.

Table 4
Some popularly used measures of market concentration and their properties

Measure	Functional form	Range	Typical features
CR_K	$\sum_{j=1}^{k} s_i^{[j]}$	$0 < CR_k \leq 1$	Takes only large banks into account; arbitrary cut-off
HHI	$\sum_{i=1}^{n} s_i^2$	$1/n \leq HHI \leq 1$	Considers all banks; sensitive to entrance of new banks
HTI	$1/\left(2\sum_{i=1}^{n} i s^{(i)} - 1\right)$	$0 < HTI \leq 1$	Emphasis on absolute number of banks
Entropy	$-\sum_{i=1}^{n} s_i \log_2 (s_i)$	$0 \leq E \leq \log_2 (n)$	Based on expected information content of a distribution

Note: Here $s^{[i]}$ and $s^{(i)}$ are the *i*-th highest and *i*-th lowest market share respectively.

In this study, we examine the possible impact on five popularly used market concentration measures: (i) 1-bank concentration ratio, (ii) 3-bank concentration ratio, (iii) 10-bank concentration ratio, (iv) HHI and (v) entropy measure. Table 5 juxtaposes these figures for the three data sets. We ignore the rank-

[8] For example, the survey of Bikker and Haaf (2001) demonstrates that for 20 countries the rankings of the k-bank concentration ratios and the HHI are strongly correlated. Similarly, the study of Bhattacharya and Das (2003) in the context of India reveals that movements in popularly used concentration measures over time for a single country could be highly correlated.

based measures because reconciling the differences in the entire rank structure in a sample and in a population could be a difficult and complicated task.

Table 5 reveals that the estimated concentration measures from the BankScope data are not far from the corresponding population figures, though they tend to overestimate them.[9] The extent of overestimation in India is clearly more if RRBs are considered as a part of the population. However, although, in absolute terms, the deviations among the three sets of values appear to be close, they need to be interpreted with caution. For example, while a merger involving two small banks (ie, in the left tail in Figures 1–3) based on the sample values may not differ much from that in the population, that involving two large banks in the right tail possibly could. It may be noted that theoretical evidence from count data models suggests the presence of an upward bias in the HHI (Hall (2000)). In the case of the Gini ratio, which closely resembles variance-based measures like the HHI, the extent of small-sample bias has been identified and appropriate correction factors have been suggested (Deltas (2003)). Our empirical results, therefore, appear to be consistent with theory.

Table 5
Comparison of market concentration measures

Market concentration measure	BankScope	Population (with RRBs)	Population (without RRBs)
1-bank CR	0.2638	0.2348	0.2437
3-bank CR	0.3725	0.3314	0.3441
10-bank CR	0.6182	0.5508	0.5719
HHI	0.0898	0.0719	0.0775
Entropy measure	4.6300	5.2558	4.9413

4. Can the selectivity bias be reduced or eradicated from the sample?

This section examines the implication of possible selectivity bias on a few popularly used market concentration measures and demonstrates ways to correct it with the help of some additional information. Earlier, in the case of HHI, the problem of its determination under incomplete information was addressed by researchers (Nauenberg et al (1997)). Nauenberg et al (1997) examine a condition where only the market shares of a few top firms are known. By using tools from combinatorics, they specified a probability model and suggested a simulation scheme to determine the population HHI. In this paper, an alternative scheme, which could be applied not only to the HHI but also on a few other concentration measures, is suggested. To do that, it is assumed that the coverage ratio with respect to the size variable is known. In the case of total assets, it implies knowledge of the figure for total assets of the banking sector as a whole in an economy. The information relating to this aspect is generally available to the central bank for policy purposes and, in most cases, publicly disseminated in a routine manner.[10] It may also be noted that knowledge of the output and market shares of the top few firms, as in Nauenberg et al (1997), implicitly assumes knowledge of the coverage ratio with respect to the size variable. It is shown here that if the coverage ratio is sufficiently large, concentration measures pertaining to the population could be approximated from the sample with the help of appropriate correction factors that are functions of the coverage ratio. Thus the simulation schemes – as suggested in Nauenberg et al (1997) for HHI – would be redundant if the coverage ratio is sufficiently large.

[9] Earlier, Corvoisier and Gropp (2001) also suspected that, due to selectivity bias in the BankScope data, the sample measure may understate the degree of concentration in some countries.

[10] For example, in the context of the euro area, Corvoisier and Gropp (2001) observed that the coverage of total assets of banks in BankScope is not complete and obtained figures for them from alternative OECD publications.

Let there be (m+n) banks in an economy. Let $a_1, a_2, ..., a_m, a_{m+1}, ..., a_{m+n}$ be the total assets of these banks. Without loss of generality, let us assume that the first m banks are included in the sample and the rest of the n banks are ignored. Also, let the total assets of the m included banks in the sample be A_S, the total assets of the n excluded banks be A_I and the total asset in the population be A_P. Clearly, $A_P = A_S + A_I$. Let the coverage ratio be denoted as C, where $C = (A_S / A_P)$.

The k-bank concentration measures

We shall first demonstrate the implication for k-bank concentration measures for k=1. Let the asset sizes of the largest banks in the sample and in the population be $a_S^{[1]}$ and $a_P^{[1]}$ respectively. Then, the sample and population 1-Bank Concentration Measures are ($a_S^{[1]} / A_S$) and ($a_P^{[1]} / A_P$) respectively. Let us now consider two cases:

Case 1: It is known that $a_S^{[1]} = a_P^{[1]}$, ie the largest bank in the population is included in the sample. In this case, it is easy to show that the population 1-Bank Concentration Measure could be obtained by simply multiplying the corresponding sample measure by C.

Case 2: It is known that $a_S^{[1]} \neq a_P^{[1]}$, ie the largest bank in the population is not included in the sample. An estimation of the population 1-Bank Concentration Measure is not easy in this case.

For general k-bank concentration measures, similar results hold. If it is known that the top k banks are included in the sample, knowledge of the coverage ratio alone would be sufficient to glean the population values. However, if data on some of the top banks are not available in the sample, estimation would be difficult. In many cases, however, the top few banks in an economy are well known, and if the selectivity bias is in favour of choosing larger banks, it is likely that they would be included in the sample. Alternatively, however, if one considers the dual problem of measuring market concentration from the lower end, ie based on the shares of k smallest banks in the total assets, the estimate from the sample possibly could not be corrected even after multiplying with the coverage ratio because sample data are perhaps more likely to exclude the “unimportant” small banks.

The coverage ratios in India when RRBs are included and excluded are 0.889921 and 0.924003 respectively. Using this and the additional knowledge that SBI, the largest bank in India, is included in the BankScope database leads one to estimate the corresponding population measures as 0.2348 and 0.2437 respectively, exactly as reported in Table 5 as population measures.

The HHI

Let the values of HHI for the sample, the ignored part of the population and the total population be denoted by HHI_S, HHI_I and HHI_P respectively. For HHI, an additive decomposition of concentration across the included and the excluded group of banks reveals that:

$$(2) \qquad HHI_P = \left(\frac{A_S}{A_P}\right)^2 HHI_S + \left(\frac{A_I}{A_p}\right)^2 HHI_I \quad = C^2\ HHI_S + (1-C)^2\ HHI_I$$

From equation (2), it is clear that as HHI_I is a bounded function taking values in [1/n, 1], one could adequately approximate HHI_P from the sample if the coverage ratio is high. This is because if C is close to unity, the second term on the right-hand side of equation (2) would be close to zero. Thus, the first term alone could be used as an estimate of population HHI.

The correction factors (C^2) estimated from the BankScope database turn out to be 0.791959 when RRBs are included and 0.853782 when RRBs are excluded, yielding estimates of HHI_P in the two cases as 0.0711 and 0.0767 respectively. A comparison with the population values in Table 5 reveals that the estimated values differ from those in the population only in the fourth decimal place. This approximation of HHI is likely to be sufficiently close to the population values, at least to the extent desired by the policymakers.

The entropy measures

Let the sample entropy, population entropy and entropy within the neglected group of banks be denoted by E_S, E_P *and* E_I respectively. It can be shown that:

$$E_P = C E_S + (1 - C) E_I - C \log_2 (C) - (1 - C) \log_2 (1 - C) \quad (3)$$

In equation (3), total entropy is decomposed into within-group and between-group entropy. Also, it is observed that, with the accurate knowledge of *C*, it is possible to obtain accurate values of the first, third and fourth terms on the right-hand side of equation (3). How good this approximation would be for E_P would depend on a few factors. It is well known that the maximum value of the entropy measure with *n* points of positive probability mass is $\log_2(n)$. Thus, if *n* is large and the value of the entropy measure corresponding to the *n* ignored banks is high, the approximation may not be sufficiently close to the population figure even with a reasonable coverage ratio of 90%. Of course, a ceteris paribus increase in *C* would increase accuracy of the estimate further. However, to increase accuracy, here it is important to increase both the coverage in terms of the size variable and the coverage in terms of the absolute number of entities. The Indian case is a relevant example to establish this point. The Indian banking market contains a large number of small RRBs and moderately sized FBs, and entropies within these groups are relatively high. Therefore, approximations through the first three terms yield the values of 4.62 and 4.67 respectively including and excluding RRBs, considerably diverging from the corresponding population values reported in Table 5.

The above results highlight the importance of knowledge of the coverage ratio. An implication of the above results is that if the coverage ratio is known and is reasonably high, then HHI is probably the most suitable tool for measuring population market concentration from sample data, despite some of its theoretical drawbacks. The paper indicates that use of such sample-based measures in cross-country, time series, or panel studies would require good degree of caution. Some of the earlier researchers like Corvoisier and Gropp (2001) were aware of the problem. However, to tackle it, they eliminated some of the banks from the BankScope data and restricted their analysis to a common set of banks. Results in this paper reveal that this “medicine” is likely to aggravate the “disease”. Instead, the use of relevant correction factors based on coverage ratios is suggested. It may be noted that the correction factors for obtaining good estimates pertaining to the population could be complex and non-linear functions of the coverage ratio. In empirical studies, it is tempting to replace the sample estimates as proxies for population measures. However, as coverage ratios in a sample may vary across countries or time, applying the sample-based measures directly for policy purpose or including in regression equations may lead to wrong conclusions. For example, if the coverage of developed economies in BankScope is good and that of developing or emerging economies poor, the upward bias in the measures would be greater for the developing economies!

5. Conclusion

The BankScope database is undoubtedly a valuable database. It is specifically useful for investors, who can not only examine the balance sheets and profit and loss accounts of individual financial entities for a number of years, but can also compare those figures with the majority of the peers and possible competitors. Barring a few minor discrepancies, this paper found that the values reported in the database are consistent with those reported in the primary sources. The discrepancies could be due to the maintenance of a uniform accounting convention in a cross-country database like BankScope.

Unfortunately, the same conclusion may not hold when one attempts to use the BankScope database for cross-country studies that need macro-level estimates for specific characteristics of the banking sector in different countries. The disseminated population data for India pointed out the presence of a strong selectivity bias in BankScope data. It was found that the BankScope database almost totally excluded Regional Rural Banks and Foreign Banks in India. As the banking market in India appears to be segmented, this selectivity bias seriously affected measurements of the moments of the distribution of the log of asset values. In particular, it might lead users of the data to conclude that the Indian banking market is unimodal when the population data actually reveal a bimodal pattern. Kolmogorov-Smironov tests revealed that neither the distribution of the log of asset values nor that of market shares from BankScope data is likely to be the same as the population distribution, irrespective of whether Regional Rural Banks in India are included or excluded from the population.

Despite these limitations, it was, however, demonstrated that a few popularly used market concentration measures – especially the HHI – could be estimated from BankScope data accurately, provided the coverage ratio with respect to the size variable is adequate and is known. For k-bank

Concentration Measures, it was shown that if the top k banks in the population were also present in the sample, the population measure could be estimated without any error using the coverage ratio. Empirical work using Indian banking data suggested that coverage of about 90% with respect to the size variable could be adequate for approximating population HHI, while for entropy measures it indicated that more coverage – with respect to both the size variable and the number of financial entities – might be required.

This study thus warns against hasty and superficial use of BankScope data in the context of country-specific or cross-country studies. Ideally, one should use the data on the entire population of banks in such studies. As these data may not be readily available, the dependence on samples is perhaps unavoidable. However, it is more important to have a "representative" sample than to increase coverage arbitrarily and in a haphazard manner. When inferences are drawn from these data as in BankScope, they may suffer from implicit selectivity bias, the characterisation of which is not an easy task from the sample. Though a few popularly used measures could be approximated accurately with a little additional information that is generally available in the public domain, even in such cases the correction factors for these measures might not be obvious and could be non-linear functions of the coverage ratio. Specification and clear articulation of the sampling design, the inclusion policy and the extent of country-specific coverage with respect to different size variables are therefore absolutely essential in any cross-country database.

References

Barth, J R, R D Brumbaugh Jr and J A Wilcox (2000): "The repeal of Glass-Stegall and the advent of broad banking", *Journal of Economic Perspectives*, 14, pp 191–204.

Bhattacharya, K and A Das (2003): "Dynamics of market structure and competitiveness of the banking sector in India and its impact on outputs and prices of banking services", Reserve Bank of India, [mimeo].

Bikker, J A and K Haaf (2001): "Measures of competition and concentration: a review of literature", De Nederlandsche Bank, Amsterdam.

Corvoisier, S and R Gropp (2001): "Bank concentration and retail interest rates", *Working Paper* no 72, European Central Bank.

Cunningham, A (2001): "Assessing the stability of emerging market economies' banking systems", *Financial Stability Review*, 11, December, pp 187–92.

De Bandt, O and E P Davis (1999): "A cross-country comparison of market structures in European banking", *Working Paper* no 7, European Central Bank.

Deltas, G (2003): "The small-sample bias of the Gini Coefficient: results and implications for empirical research", *Review of Economics and Statistics*, 85, pp 226–34.

Demirgüç-Kunt, A and E Detragiache (1998): "Financial liberalization and financial fragility", *IMF Working Paper* no 98/83.

Ennis, H M (2001): "On the size distribution of banks", *Federal Reserve Bank of Richmond Economic Quarterly*, 87/4, Fall, pp 1–25.

Hall, B H (2000): "A note on the bias in the Herfindahl based on count data", UC Berkley and Nuffeld College, Oxford, [mimeo].

Marfels, C (1971): "Absolute and relative measures of concentration reconsidered", *Kyklos*, 24, pp 753–66.

Mathieson, D J and J Roldos (2001): "Foreign banks in emerging markets" in R Litan, P Masson and M Pomerleano (eds), *Open doors: foreign participation in financial systems in developing countries*, Brookings Institution Press.

Nauenberg, E, K Basu and H Chand (1997): "Hirschman-Herfindahl Index determination under incomplete information", *Applied Economics Letters*, 4, pp 639–42.

Neumann, M (2001): *Competition policy: history, theory and practice*, Edward Elgar, Cheltenham, United Kingdom.

Reserve Bank of India (2001a): *Report on trend and progress in banking in India 2000-2001*.

—— (2001): *Statistical tables relating to banks in India 2000-2001*.

Singh, A (2002): "Competition and competition policy in emerging markets: international and developmental dimensions", *Working Paper* no 246, ESRC Centre for Business Research, University of Cambridge.

Recent BIS Working Papers

No	Title	Author
132 July 2003	Developing country economic structure and the pricing of syndicated credits	Yener Altunbaş and Blaise Gadanecz
131 March 2003	Optimal supervisory policies and depositor-preference laws	Henri Pagès and João A C Santos
130 February 2003	Living with flexible exchange rates: issues and recent experience in inflation targeting emerging market economies	Corrinne Ho and Robert N McCauley
129 February 2003	Are credit ratings procyclical?	Jeffery D Amato and Craig H Furfine
128 February 2003	Towards a macroprudential framework for financial supervision and regulation?	Claudio Borio
127 January 2003	A tale of two perspectives: old or new challenges for monetary policy?	Claudio Borio, William English and Andrew Filardo
126 January 2003	A survey of cyclical effects in credit risk measurement models	Linda Allen and Anthony Saunders
125 January 2003	The institutional memory hypothesis and the procyclicality of bank lending behaviour	Allen N Berger and Gregory F Udell
124 January 2003	Credit constraints, financial liberalisation and twin crises	Haibin Zhu
123 January 2003	Communication and monetary policy	Jeffery D Amato, Stephen Morris and Hyun Song Shin
122 January 2003	Positive feedback trading under stress: Evidence from the US Treasury securities market	Benjamin H Cohen and Hyun Song Shin
121 November 2002	Implications of habit formation for optimal monetary policy	Jeffery D Amato and Thomas Laubach
120 October 2002	Changes in market functioning and central bank policy: an overview of the issues	Marvin J Barth III, Eli M Remolona and Philip D Wooldridge
119 September 2002	A VAR analysis of the effects of monetary policy in East Asia	Ben S C Fung
118 September 2002	Should banks be diversified? Evidence from individual bank loan portfolios	Viral V Acharya, Iftekhar Hasan and Anthony Saunders
117 September 2002	Internal rating, the business cycle and capital requirements: some evidence from an emerging market economy	Miguel A Segoviano and Philip Lowe
116 September 2002	Credit risk measurement and procyclicality	Philip Lowe
115 August 2002	China's asset management corporations	Guonan Ma and Ben S C Fung
114 July 2002	Asset prices, financial and monetary stability: exploring the nexus	Claudio Borio and Philip Lowe
113 July 2002	The link between default and recovery rates: effects on the procyclicality of regulatory capital ratios	Edward I Altman, Andrea Resti and Andrea Sironi
112 June 2002	Determinants of international bank lending to emerging market countries	Serge Jeanneau and Marian Micu
111 April 2002	Output trends and Okun's law	Gert Schnabel