



The analysis of data errors in financial information databases: New evidence from the Korean financial markets

**Authors:**

Hyung-Chan Jung¹ 
Hyun-Jung Nam² 

Affiliations:

¹Division of Business Administration, Pukyong National University, Republic of Korea

²Graduate School of International Studies, Dong-A University, Republic of Korea

Corresponding author:

Hyun-Jung Nam,
hjn Timer@daa.ac.kr

Dates:

Received: 28 Mar. 2017

Accepted: 26 Aug. 2017

Published: 27 June 2018

How to cite this article:

Jung, H-C. & Nam, H-J., 2018, 'The analysis of data errors in financial information databases: New evidence from the Korean financial markets', *South African Journal of Business Management* 49(1), a185. <https://doi.org/10.4102/sajbm.v49i1.185>

Copyright:

© 2018. The Authors.
Licensee: AOSIS. This work is licensed under the Creative Commons Attribution License.

Background: As financial professionals including policy-makers tend to base decisions on research performed using large machine-readable financial databases, the accuracy of the financial data provided by database companies has a direct impact on the quality of their decisions.

Objectives: The objective of this study was to examine data errors in the DataGuide and KisValue databases which are both primary sources of stock prices and return data for Korea Exchange securities in Korea. This article also discussed the methodological implications of erroneous data on monthly stock returns in empirical studies on Korean financial markets.

Methods: A cross-checking technique was used in this study.

Results: The results suggest that there are material discrepancies between the DataGuide and KisValue databases in monthly stock returns, most of which are attributable to the mishandling of split events and of missing values. The results also indicate that DataGuide provides a more reliable service than KisValue in terms of monthly stock returns.

Conclusion: The results show that extreme monthly returns resulting from serious data errors in the DataGuide and KisValue databases may be enough to sharply change the properties of monthly stock return distributions and to over- or underestimate long-term abnormal stock returns.

Introduction

This article examines data errors in the DataGuide and KisValue databases, which are both primary sources of the stock prices and return data for Korea Exchange (KRX) securities in Korea, by using a cross-checking technique. We also discuss the methodological implications of erroneous data for monthly stock returns in empirical studies on Korean financial markets. In order to do that, we match and compare monthly stock returns for shares listed on the KRX Securities Market that are available in both the DataGuide and KisValue databases and analyse the accuracy of the data as well as the source of errors in each database, covering 15 years from January 2000 to December 2014.

As financial professionals, including policy-makers, tend to base decisions on research performed by using large machine-readable financial databases, the accuracy of the financial data provided by database companies has a direct impact on the quality of their decisions. That is, if the financial data employed by decision-makers are accurate, they can make high-quality decisions that are appropriate for research (Winkler, Kuklinski & Moser 2015). On the contrary, the presence of erroneous data might distort their decisions and seriously damage the interests of individuals, firms as well as the entire economy.

In fact, previous studies have suggested that data errors can be a serious problem in computerised financial databases such as Center for Research in Security Prices (CRSP), Compustat and Value Line (Bennin 1980; Chychyla & Kogan 2015; Kinney & Swanson 1992; Rosenberg & Houglet 1974; San Miguel 1977; Tallapally, Luehlfling & Motha 2011; Yang, Vasarhelyi & Liu 2003). Rosenberg and Houglet (1974) were the first to examine the error rates in the CRSP and Compustat databases and their methodological implications by using a cross-checking technique. They matched and compared monthly price relatives for 844 industrials from January 1963 to June 1968, and for 97 utilities from March 1962 to June 1968. They found that 1060 (2.99%) out of 35 357 industrial price relatives and 142 (2.39%) out of 5939 utility price relatives were erroneous. They also found that four of the 34 discrepancies, greater than 0.05 between CRSP and Compustat databases, were because of CRSP errors, and 30 because of Compustat errors, suggesting that CRSP is a better

Read online:

Scan this QR code with your smart phone or mobile device to read online.

database for monthly price relatives. Bennin (1980) conducted a follow-up study to that of Rosenberg and Houglet (1974), using updated Compustat and CRSP databases. In his study, the monthly return with dividends for each New York Stock Exchange (NYSE) company on Compustat was matched against its CRSP monthly return. Even though using monthly returns, including dividends, added the possibility of errors caused by incorrect dividend information, the overall error rate turned out to be only one-third of the rate reported by Rosenberg and Houglet (1974). Bennin (1980) suggests that Compustat errors dropped markedly after 1970, indicating that data collection improved, following the Rosenberg and Houglet (1974) study.

Since the early 1970s, when Rosenberg and Houglet first reported a few large errors in CRSP and Compustat data on monthly price relatives, researchers have also investigated potential data problems in accounting databases such as Compustat, Value Line and EDGAR Online (Chychyla & Kogan 2015; Kinney & Swanson 1992; San Miguel 1977; Tallapally et al. 2011; Yang et al. 2003). Yang et al. (2003) examined the accuracy of seven frequently used accounting variables in Compustat and Value Line databases during the 11 years from 1971 to 1981. Out of 10 353 comparisons, 1284 (12.4%) were, with discrepancies, greater than 1%. In order to identify the cause of the discrepancies, they compared a subsample of the 1981 data to the original financial statements. They found that the sources of discrepancies could be classified into the following two types: (1) explainable definitional differences (i.e. foreign currency differences, industry and definitional factors); and (2) unexplained differences (i.e. non-disclosed coding rule differences and coding errors). Chychyla and Kogan (2015) conducted the first large-scale comparison of Compustat and 10-K data by comparing 30 accounting variables for approximately 5000 companies from 2011 to 2012. The results showed that the values, reported in Compustat, were significantly different from those reported in 10-K filings. They also showed that the amount and magnitude of the original data alterations, introduced by Compustat, depended on the type of accounting variable and firm characteristics such as industry and size.

Until recently, many studies in the United States have documented that (1) there exist erroneous data in well known commercial financial databases such as CRSP, Compustat and Value Line; and (2) a few serious errors in these databases could adversely affect research and decision-making. In Korea, however, it is very hard to find studies that examine data errors in financial information databases, except for those by Oh and Lee (2007), and Nam (2016). Oh and Lee (2007) matched and compared the data values of five selected accounting variables (i.e. Total Assets, Total Liabilities, Sales, Net Income and Operating Cash Flow) provided by the three prominent data aggregators in Korea: DataGuide, KisValue and TS2000. Out of 3500 observations, they found 818 (23.4%) mismatches between the three accounting databases for the period of 2000–2006. The largest source of mismatches was differences in the policies of data aggregators to restate past data that had subsequently been changed by the company

because of an accounting restatement. They suggested that the results of an accounting empirical study should depend on the database used for the study. Nam (2016) investigated data discrepancies between the DataGuide and TS2000 databases, comparing the values of 10 selected accounting variables for the years 2011 through 2013 in the two databases. They found that 1194 (5.88%) of 20 310 observations were discrepant and there were statistically significant differences in five variables out of 10.

In comparison to United States studies, few studies in Korea have examined the quality of accounting data in popular financial databases. The stock returns data for KRX securities, provided by data aggregators, have never been verified by using a cross-checking technique, even though they are frequently used in empirical studies in both corporate finance and investment (Jung 2010; Kho & Park 2000; Kim, Kim & Shin 2012; Lee & Cho 2014). Furthermore, none of the existing studies attempted to identify the most reliable database in terms of error rates in competing databases. We therefore will examine the accuracy of monthly returns for KRX securities in the DataGuide and KisValue databases for the 15 years - from January 2000 to December 2014. We chose these two databases for the study, because they are the ones most commonly used by academicians and financial professionals in Korea (Baik, Kang & Kim 2013; Baik et al. 2015; Chang & Shin 2007; Choi, Sohn & Seo 2015; Kim, Lee & Shin 2017). We also discuss the methodological implications of data errors on monthly stock returns in empirical research on Korean financial markets. In addition, we also point out which database is superior in reporting accurate monthly stock returns. Our results show that there are material discrepancies in monthly stock returns between the DataGuide and KisValue databases. Most of the errors are attributable to the mishandling of missing values and of split events (i.e. stock splits, capital reductions, rights offerings and spin-offs). Our results also indicate that DataGuide is more reliable than KisValue in terms of monthly stock returns. Finally, our results suggest that extreme erroneous returns in the DataGuide and KisValue databases may be enough to sharply affect the properties of monthly stock return distributions and to over- or underestimate long-run abnormal stock returns.

The remainder of this article is organised as follows. The 'Data and methodology' section discusses the data and methodology. The 'Analysis and results' section analyses the discrepancies between the DataGuide and KisValue databases in the data on monthly stock returns. The 'Methodological implications of data errors' section presents and discusses the primary results, and the 'Summary and conclusions' section concludes this article.

Data and methodology

Data

It is well known that DataGuide (from FnGuide, Inc.) and KisValue (from NICE Information Service) are the primary sources of both stock return data and historical accounting data in Korea. In order to examine the quality of stock

return data in these two popular databases, we matched and compared DataGuide and KisValue data on monthly returns for 729 KRX listed securities for a period of 15 years - from January 2000 to December 2014. In the study by Rosenberg and Houglet (1974), the comparison of monthly price relatives data between the CRSP and Compustat databases is clearly asymmetric, because monthly stock prices are among the least important of the data in the Compustat database, whereas they are the primary content of the CRSP database. However, when comparing DataGuide and KisValue data on monthly returns in this study, we expected no asymmetry that would lend an advantage to either database, because both DataGuide and KisValue are universal financial databases in Korea.

To be included in the sample, KRX securities had to meet the following criteria:

1. The data on the monthly returns of common shares should be available in both DataGuide and KisValue databases.
2. The data on daily and monthly stock prices should be available at the KRX website (www.krx.co.kr).
3. The annual reports and major disclosure information of the companies should be available at DART (dart.fss.or.kr), the electronic disclosure system of the Financial Supervisory Service or KIND (kind.krx.co.kr), the electronic disclosure system of KRX.

Applying these criteria, we compared a total of 109 260 firm-month data on monthly stock returns between the DataGuide and KisValue databases.

Cross-checking technique

For this study, we matched and compared the monthly returns data for KRX listed securities between the DataGuide and KisValue databases by using a cross-checking technique. Rosenberg and Houglet (1974) were first to use cross-checking to compare monthly price relatives for NYSE listed stocks available on both CRSP and Compustat databases. Cross-checking is defined as a method of identifying discrepancies by matching and comparing data on selected financial variables, provided by two different indexed databases, during the same period (year, month or day).

According to Rosenberg and Houglet (1974), errors in the two databases that are truly independent, which do not reflect an error in a source used by both, will be detected with a very high degree of probability in a comparison. They also argue that such a comparison is not only the most effective way to screen for data errors, but also the least expensive. However, cross-checking is not always a perfect tool, because it cannot detect erroneous data if the same error occurs in both databases at the same time.

Calculating monthly stock returns

A monthly stock return is the change in the total value of an investment in a stock after a month per dollar of initial investment. In this study, monthly stock-return

means a monthly stock-return without dividends. A monthly stock-return is therefore calculated, as shown in equation 1.

$$r(t) = \frac{p(t)f(t) - p(t-1)}{p(t-1)} \quad [\text{Eqn 1}]$$

where

t = a holding period

$t - 1$ = time of last available price

$r(t)$ = return on purchase at $t - 1$, sale at t

$p(t)$ = last sales price or closing bid and ask average at time t

$f(t)$ = factor to adjust price in month t

$p(t - 1)$ = last sale price or closing bid and ask average at time $t - 1$.

In equation (1), $(t - 1)$ is usually 1 month before t , but it can be up to 10 months before t if there are no valid prices in the interim. The factor to adjust price in month t , $f(t)$, is one plus the number of additional shares per old share issued for stock splits. For example, if a 2-for-1 stock split is the only distribution event in a specific month, the factor to adjust price in month t is 2.

Analysis and results

The comparison of DataGuide and KisValue databases

To examine the quality of stock returns data in commercial financial databases in Korea, we matched and compared a total of 109 260 firm-month data on monthly stock returns between the DataGuide and KisValue databases. The data on monthly stock returns were downloaded from the websites of the two databases on the Internet (<http://www.dataguide.co.kr/> and <http://www.kisvalue.com/>) during 30 June 2015 to 3 July 2015. Table 1 reports the numbers and percentages of discrepancies between monthly stock-return data in the DataGuide and KisValue databases by the level of the discrepancy.

As shown in Table 1, we matched and compared a total of 109 260 monthly stock returns between the DataGuide and KisValue databases. Out of 109 260 comparisons, we found 2563 (2.35%) to be discrepant, including 381 (0.35%) that differed by more than 1%, 58 (0.05%) that differed by more

TABLE 1: The comparison of monthly stock returns between the DataGuide and KisValue databases.

Monthly stock returns matched	Number	Percentage
	109 260	100.00
Level of discrepancy:		
More than 20%	24	0.02
More than 5%, but less than 20%	58	0.05
More than 1%, but less than 5%	381	0.35
Less than 1%	2100	1.93
Total	2563	2.35

This table reports the numbers and percentages of discrepancies between monthly stock returns data in the DataGuide and KisValue databases by the level of discrepancy. We compared data on monthly returns for 729 KRX listed securities for a period of 15 years - from January 2000 to December 2014 - between the DataGuide and KisValue databases. In total, we compared 109,260 monthly stock returns between the two databases. The data on monthly stock returns were downloaded from the websites of these two databases on the Internet (<http://www.dataguide.co.kr/> & <http://www.kisvalue.com/>) between 30 June 2015 and 03 July 2015.

than 5% and 24 (0.02%) that differed by more than 20%. That is, the number (percentage) of discrepancies greater than 1% between the DataGuide and KisValue databases was 463 (0.42%) which is much lower than the counterpart 1060 (3.00%) for the industrials between the CRSP and Compustat databases in the study by Rosenberg and Houglet (1974). Further, the number (percentage) of discrepancies greater than 5% between the DataGuide and KisValue databases was only 82 (0.07%) which is also lower than the counterpart 467 (0.27%) between the CRSP and Compustat databases in Bennin (1980). These results suggest that the DataGuide and KisValue databases, which are commonly used by academicians and financial professionals in Korea, provide relatively accurate data on monthly stock returns in terms of the discrepancy rate.

Sources of errors

Even though the discrepancies of monthly stock returns data between the DataGuide and KisValue databases could stem from a variety of sources, we categorised them into the following four types of errors: (1) mishandling of split events; (2) mishandling of missing returns; (3) misspecification of month-end dates and (4) unexplainable errors.

Mishandling of split events

If you own a stock that undergoes a split event, such as stock split, stock dividend, capital reduction, right offerings and spin-off, you should use the split-adjusted price when calculating stock-return. In other words, the last sale price, $p(t)$, should be adjusted for a specific split event by using an appropriate adjustment factor in a certain month, $f(t)$, as suggested in equation (1). Otherwise, the error in handling the split event leads to a serious erroneous monthly stock-return in proportion to the split ratio, which is almost always large, ranging from 2.0 to 72.0 in the sample in this study. Let's take a couple of examples to see how the mishandling of split events could result in serious erroneous data on monthly stock returns in the DataGuide and KisValue databases.

As an example of the consequence of the mishandling of split events, we can present the case of calculating monthly returns on the common shares of LG Chemicals in April 2009. On 19 December 2008, LG Chemicals announced that it would spin-off the industrial material business, now called LG Houses, on 01 April 2009. The old shareholders of LG Chemicals received 0.12 common shares of the newly established LG Houses as well as 0.88 common shares of existing LG Chemicals in exchange for one old common share of LG Chemicals. Because of the spin-off procedure, LG Chemicals common shares were suspended after they were traded at the closing price of 90 000 KRW on 27 March 2009 until the suspension was lifted on 20 April 2009. When the new common shares of LG Chemicals and LG Houses were relisted on the KOSPI Market of KRX on 20 April 2009, trading resumed at the beginning price of 128 000 KRW for LG Chemicals common shares. On the last trading day of April 2009, the share prices of LG Chemicals and LG Houses

closed at 141 500 KRW and 115 000 KRW, respectively. In this example, the adjustment factor, $f(t)$, for the last sale price of LG Chemicals common stock can be estimated, as shown in equation 2:

$$f(t) = \frac{[141,500(0.88) + 115,000(0.12)]}{141,500} = 0.9776 \quad [\text{Eqn 2}]$$

Applying the adjustment factor estimated by equation (2), we can calculate the monthly return on LG Chemicals common stock in April 2009, as suggested in equation 3:

$$r(t) = \frac{[141,500(0.9776) - 90,000]}{90,000} = 0.5368 \quad [\text{Eqn 3}]$$

However, DataGuide made a fatal error in calculating the monthly return on LG Chemicals common stock in April 2009, because it failed to take into account the effect of the spin-off event on the month-end price, $p(t)$. DataGuide calculated the monthly return in a manner that completely ignored the effect of the spin-off event on the total value of an investment in LG Chemicals common shares, and simply used the purchase and sales price of LG Chemicals shares, as shown in equation 4, despite the fact that old shareholders of LG Chemicals received both 0.12 shares of LG Houses common stock and 0.88 shares of LG Chemicals common stock for one share of LG Chemicals common stock through the spin-off event in April 2009.

$$\text{DataGuide: } r(t) = \frac{[141,500 - 128,000]}{128,000} = 0.1050 \quad [\text{Eqn 4}]$$

In addition, because DataGuide used the beginning price of LG Chemicals shares on 20 April 2009 as the last sale price at $t-1$, $p(t-1)$, the monthly return calculated by DataGuide (0.1050) was underestimated in comparison with the monthly return (0.5368) which appropriately reflected the effect of the spin-off event using the adjustment factor, $f(t)$, as shown in equation 3.

Meanwhile, the error that KisValue made in calculating the monthly return on LG Chemicals common stock was quite similar to that of DataGuide in that it also failed to use the month-end price, $p(t)$, adjusted to the spin-off event, as shown in equation 5.

$$\text{KisValue: } r(t) = \frac{[141,500 - 90,000]}{90,000} = 0.5722 \quad [\text{Eqn 5}]$$

The only difference between the two databases is that KisValue employed the last sale price (90 000 KRW) on 27 March 2009 when KRX suspended the trading of LG Chemicals shares as a purchase price at $t-1$, $p(t-1)$, while DataGuide used its beginning price (128 000 KRW) on 20 April 2009 when the suspension was lifted. Thus, this example demonstrates that a serious data error in DataGuide and KisValue databases could occur because of an error in handling the spin-off event.

Another example of the consequence of mishandling split events is the case of Schnell Biopharmaceuticals in May 2009.

Schnell Biopharmaceuticals conducted capital reduction without refund by consolidating 10 shares of common stock into one share on 27 May 2009. However, both DataGuide and KisValue databases failed to reflect the effect of the 1-for-10 reverse stock split on the month-end price, $p(t)$, resulting in a discrepancy greater than 2100% between the two databases. These two examples from LG Chemicals and Schnell Biopharmaceuticals clearly show that errors in handling split events, such as spin-offs and reverse stock splits, cause serious data errors in monthly stock returns in the DataGuide and KisValue databases.

Mishandling of missing returns

If you have no valid last sale price at either month t or month $t - 1$, you have to report the month t return as a missing value. Otherwise, the mishandling of missing returns might have a significant effect on the conclusions drawn from monthly returns data. The CRSP database assigns a series of special missing-return codes such as -66.0, -77.0, -88.0 and -99.0 which specify the reason why a return is missing. For example, a missing-return code of -99.0 in the CRSP database indicates that the return is missing because of a missing price at time t , usually because of suspension in trading or trading on an unknown exchange (CRSP 2012:40).¹ In contrast, neither the DataGuide nor the KisValue database has any specific policy for handling missing returns, which can occur for many reasons. Instead, both databases replace all missing returns with '0' or with extreme values, even when KRX suspends the trading of a share for more than a year, including the current month t .

An unexpected critical error resulting from the mishandling of missing returns is illustrated by the monthly stock-return on the common stock of Chinhung International, Inc. in March 2012. For nearly 2 months, from 24 February 2012 to 16 April 2012, KRX suspended trading in the common shares of Chinhung International Inc. because of a 1-for-10 reverse stock split for a shareholders' equity reduction implemented on 16 March 2012 and resumed trading on 17 April 2012. Because of the trading suspension, there was no trading in the common shares of Chinhung International Inc. on KRX for the whole month of March 2012. Therefore, the monthly return on the common shares of Chinhung International Inc. in March 2012 should be missing. As a result of an error in handling the missing-return, however, a large discrepancy, as high as 900%, between DataGuide and KisValue databases was generated. Firstly, DataGuide calculated a 0% return in March 2012, considering the effect of the reverse stock split implemented on 16 March 2012, as shown in equation 6.

$$\text{DataGuide: } r(t) = \frac{\left[2,300\left(\frac{1}{10}\right) - 230\right]}{230} = 0 \quad [\text{Eqn 6}]$$

In equation 6, the problem with the DataGuide calculation is that neither of the closing prices for the month-end dates in February and March 2012, which were 230 KRW and

2300 KRW, respectively, is valid; both prices should have been omitted because of the suspension in trading from 24 February 2012 to 16 April 2012. In fact, the closing price for the month-end date in February, 230 KRW, is arbitrarily assigned according to KRX's normal practice of replacing a missing price by the latest valid price of 23 February 2012, while the closing price for the month-end date in March, 2300 KRW, is simply an adjusted price to reflect the effect of the 1-for-10 reverse stock split conducted on 16 March 2012. Meanwhile, KisValue contained a much more serious error than DataGuide in that it neither handled the missing prices properly nor took into consideration the effect of the 1-for-10 reverse stock split in calculating the monthly return, as shown in equation 7.

$$\text{KisValue: } r(t) = \frac{[2,300 - 230]}{230} = 9.0 \quad [\text{Eqn 7}]$$

This example of the monthly return on the common shares of Chinhung International Inc., suggests that the mishandling of missing returns could cause material errors in calculating monthly stock returns in the DataGuide and KisValue databases.

Misspecification of month-end dates

KisValue errors in specifying the last trading day for January 2012 caused 394 erroneous monthly stock returns, as shown in Table 2. These errors arose, because KisValue mistakenly specified 30 January 2012 as the last trading day in January 2012, even though the correct date was 31 January 2012. As a result, KisValue made errors in calculating the monthly returns for some KRX securities traded in January 2012 by using the closing price of the misspecified trading day as the last sale price of the month, $p(t)$. For example, the month-end price for Youngbo Chemical common stock in December 2011 was 3190 KRW, while its closing prices on 30 January 2012 and 31 January 2012 were 4585 KRW and 5270 KRW, respectively. Nonetheless, KisValue made an error in calculating the monthly return in the month of January 2012, because KisValue used the closing price (4585 KRW) on January 30 instead of the closing price (5270 KRW) on January 31 as the month-end price of January 2012, as shown in equation 8.

$$\text{KisValue: } r(t) = \frac{[4,585 - 3,190]}{3,190} = 0.4373 \quad [\text{Eqn 8}]$$

Consequently, KisValue underestimated the monthly return on Youngbo Chemical common stock in the month of January 2012 by more than 20% in comparison with the correct monthly return in equation 9.

$$\left(0.6520 = \frac{[5,270 - 3,190]}{3,190}\right) \quad [\text{Eqn 9}]$$

Unexplainable errors

There were 16 discrepancies greater than 1% between the DataGuide and KisValue databases of which the source was obscure, as shown in Table 2. At our request, database companies could not explain a reasonable cause for these errors except as simple mistakes in calculating monthly stock returns.

1. For another example, in the case of a valid current price, $p(t)$, but no valid previous price, $p(t - 1)$, the month t return is given a missing-return code of -66.0.

TABLE 2: The distribution of discrepancies greater than 1% between DataGuide and KisValue by the source and level of the discrepancy.

Level of discrepancy	Source of discrepancy				Total
	Mishandling of split events	Mishandling of missing returns	Misspecification of month-end dates	Unexplainable errors	
More than 20%	11 (2.38%)	7 (1.51%)	4 (0.86%)	2 (0.43%)	24 (5.19%)
≥ 5% but < 20%	5 (1.08%)	-	49 (10.58%)	4 (0.86%)	58 (12.52%)
≥ 1% but < 5%	30 (6.48%)	-	341 (73.65%)	10 (2.16%)	381 (82.29%)
Total	46 (9.94%)	7 (1.51%)	394 (85.10%)	16 (3.45%)	463 (100%)

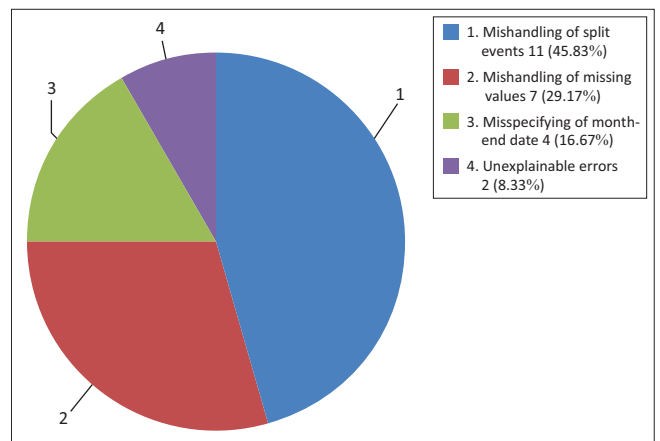
This table presents the distribution of 463 discrepancies greater than 1% between the DataGuide and KisValue databases by the source and level of discrepancy. The sources of discrepancies are categorised into the following four types of errors: (1) mishandling of split events; (2) mishandling of missing returns; (3) misspecification of month-end dates; and (4) unexplainable errors. The numbers in parentheses represent the percentages.

Table 2 shows the distribution of 463 discrepancies greater than 1% between monthly stock return data in the DataGuide and KisValue databases by the source and level of the discrepancy.

Table 2 indicates that errors in specifying month-end dates result in the highest number (percentage) of discrepancies greater than 1%, 394 (85.10%), followed by the mishandling of split events, 46 (9.94%); unexplainable errors, 16 (3.45%); and the mishandling of missing returns, 7 (1.51%). As shown in Figure 1, however, errors in handling split events have the highest number (percentage) of material discrepancies above 20%, 11 (45.83%). In contrast, the misspecification of month-end dates has only 4 (16.67%) material discrepancies greater than 20%.

In Figure 1, we find strong evidence that the mishandling of stock split events in calculating monthly stock returns is the largest source of serious discrepancies greater than 20% between the DataGuide and KisValue databases. This finding is quite similar to the results of Rosenberg and Houglet (1974), suggesting that, for the utility sample, 17 of the 34 discrepancies above 5% between the CRSP and Compustat databases were because of errors in the handling of stock splits. This type of error resulting in material discrepancies greater than 20% between the DataGuide and KisValue databases could significantly influence the results of empirical research on Korean financial markets, because they can generate erroneous extreme values or illegitimate outliers in monthly stock returns. For this reason, companies that provide databases should develop a more sophisticated algorithm to calculate the appropriate adjustment factor, taking into account the effect of split events on the stock price in order to minimise the probability of generating extreme values or outliers. In addition, they should also adopt a series of special missing-return codes, specifying the reason for missing returns as in the CRSP database to reduce large errors resulting from the mishandling of missing returns.

Table 3 presents the distribution of discrepancies that were greater than 1% between monthly stock return data in the DataGuide and KisValue, and that resulted because the databases contained errors. In Table 3, 'DataGuide and KisValue' indicates that both databases made errors in calculating monthly stock returns. As shown in Table 3, of 463 data errors resulting in discrepancies greater than 1%, 35 (= 13 + 22) resulted from DataGuide errors and 450 (= 428 + 22) from KisValue errors. Meanwhile, 19 (= 1 + 18) of 24 serious



This figure shows the distribution of 24 material discrepancies greater than 20% between the DataGuide and KisValue databases by the source of the discrepancy. The sources of discrepancies are categorised into the following four types of errors: (1) mishandling of split events, (2) mishandling of missing returns, (3) misspecification of month-end dates and (4) unexplainable errors. The numbers in parentheses represent the percentages.

FIGURE 1: The distribution of material discrepancies greater than 20% between DataGuide and KisValue by the source of discrepancy.

errors (of more than 20%) were caused by DataGuide errors and 23 (= 5 + 18) by KisValue errors. Out of 109,260 observations, the error rates in the DataGuide and KisValue databases are 0.03% and 0.41%, respectively. Hence, DataGuide seems to be a more reliable database than KisValue for monthly stock returns data. However, to ensure fairness to KisValue, KisValue can probably reduce the error rate significantly without any costly restructuring of its system for calculating monthly stock returns, because 394 of the 450 KisValue errors are simple errors that arose from misspecifying the month-end date of January 2012.

Methodological implications of data errors

The effect of errors on distributional properties of monthly stock returns

Previous studies have already demonstrated that a few large errors can have a significant impact on the distributional properties of select financial variables, including monthly stock returns (Beedles & Simkowitz 1978; Rosenberg & Houglet 1974; Yang et al. 2003). In order to examine the effect of large errors on the distributional properties of monthly stock returns, we calculate the mean, variance, skewness and kurtosis for monthly stock returns by using the two databases, DataGuide and KisValue, with different error rates. Table 4 shows four moments for monthly stock returns, including minimum and maximum values for DataGuide and KisValue databases.

TABLE 3: The distribution of data errors resulting in discrepancies greater than 1% between monthly stock returns data in the DataGuide and KisValue by databases containing errors.

Level of discrepancy	Source of errors	Database containing errors			Total
		DataGuide	KisValue	DataGuide and KisValue	
More than 20%	Mishandling of split events	-	1 (0.22%)	10 (2.16%)	11 (2.38%)
	Mishandling of missing returns	-	-	7 (1.51%)	7 (1.51%)
	Misspecification of month-end dates	-	4 (0.86%)	-	4 (0.86%)
	Unexplainable errors	1 (0.22%)	-	1 (0.22%)	2 (0.43%)
	Sub-total	1 (0.22%)	5 (1.08%)	18 (3.89%)	24 (5.18%)
More than 5%, less than 20%	Mishandling of split events	-	2 (0.43%)	3 (0.65%)	5 (1.08%)
	Mishandling of missing returns	-	-	-	-
	Misspecification of month-end dates	-	49 (10.58%)	-	49 (10.58%)
	Unexplainable errors	4 (0.86%)	-	-	4 (0.86%)
	Sub-total	4 (0.86%)	51 (11.01%)	3 (0.65%)	58 (12.52%)
More than 1%, less than 5%	Mishandling of split events	-	30 (6.48%)	-	30 (6.48%)
	Mishandling of missing returns	-	-	-	-
	Misspecification of month-end dates	-	341 (73.65%)	-	341 (73.65%)
	Unexplainable errors	8 (1.72%)	1 (0.22%)	1 (0.22%)	10 (2.16%)
	Sub-total	8 (1.72%)	372 (80.35%)	1 (0.22%)	381 (82.29%)
Total		13 (2.81%)	428 (92.44%)	22 (4.75%)	463 (100%)

This table presents the distribution of data errors resulting in discrepancies greater than 1% between monthly stock returns data in DataGuide and KisValue by databases containing errors. In the table, 'DataGuide and KisValue' indicates that both DataGuide and KisValue databases made errors in calculating monthly stock returns. The sources of discrepancies are categorised into the following four types of errors: (1) mishandling of split events; (2) mishandling of missing returns; (3) misspecification of month-end dates; and (4) unexplainable errors. The numbers in parentheses represent the percentages.

TABLE 4: The effect of large errors on distributional properties of monthly stock returns.

Distributional properties	DataGuide (A)	KisValue (B)	B/A
Sample size	109 260	109 260	-
Minimum	-0.8584	-0.9044	1.05
Maximum	7.7702	71.0000	9.14
Mean	0.0154	0.0173	1.12
Variance	0.0312	0.1069	3.43
Skewness	5.4615	126.0659	23.08
Kurtosis	123.1019	25550.4000	207.55

This table presents the distributional properties (mean, variance, skewness and kurtosis) of monthly stock returns for two databases, DataGuide and KisValue, with different error rates. In the table, 'B/A' means the ratio of distributional properties in KisValue (B) to those in the DataGuide (A) database.

As the frequency of large errors in the handling of split events and missing returns is higher for the KisValue database than for the DataGuide database, as shown in Table 3, it appears that the frequency of erroneous extreme returns might be higher for the KisValue database. The minimum and maximum values of monthly stock returns in the DataGuide database are all legitimate, while those in the KisValue database are all erroneous in Table 4. Table 4 also shows that the mean of monthly stock returns for the DataGuide database, 0.0154, is very similar to the mean for the KisValue database, 0.0173, implying that the higher frequency of extreme returns did not influence the means of monthly returns distribution. However, it turns out that higher moments for the monthly stock returns distributions are more significantly affected, as is consistent with the results of Rosenberg and Houglet (1974). In other words, the higher frequency of extreme returns resulting from errors in handling split events and missing returns, affects the means by factors of less than 1/100, the variances by factors as great as 3, the skewness by factors as great as 23 and the kurtosis by factors as great as 207. Another interesting feature of these results is that the normality assumption for monthly stock returns could not be accepted because of the skewness of greater than five and the kurtosis of greater than 100 which is consistent with the results of Koo (1998).

The effect of errors on long-term stock performance

Another consequence of a few erroneous extreme returns is the over- or underestimation of the long-term stock performance of the individual securities. Table 5 shows the effect of large errors on the long-term abnormal stock returns for a sample of 24 firm-months for which there were material discrepancies greater than 20% between the DataGuide and KisValue databases. In order to examine the effect of large errors on the long-term stock performance for this sample, Table 5 contrasts the long-term abnormal stock returns estimated, using erroneous monthly returns data with those estimated use of corrected monthly returns data found by correcting 24 large errors detected in the DataGuide and/or KisValue databases. If both the DataGuide and KisValue databases made errors in the same month for a certain stock, a long-term abnormal stock-return is calculated by using more severely erroneous data.

In order to measure the long-term stock performance in Table 5, we calculated a 36-month cumulative abnormal return (CAR) and buy-and-hold abnormal return (BHAR) using the KOSPI equally weighted market index as a return benchmark.² We defined the event month ($t = 0$) as 1 month before a large error occurred and calculated CAR and BHAR for a sample firm over a 36-month window, starting from 1 month after the event month and ending 36 months after the event month. Thus, the 36-month CAR and BHAR are defined as equations 10 and 11, respectively.

$$CAR_j = \sum_{t=1}^{36} [r_{jt} - r_{Et}] \quad [\text{Eqn 10}]$$

2. Jung (2007) suggests that for detecting long-run abnormal stock returns, BHARs calculated using the book-to-market/size-matched control firm method yield well-specified as well as the most statistically reliable test statistics in the Korean stock market. For simplicity's sake, however, we use the KOSPI equally weighted market index as a return benchmark in calculating a 36-month CAR and BHAR.

TABLE 5: The effect of large errors on long-term abnormal stock returns.

Company	Date of error	Long-term stock performance for erroneous data		Source containing error	Long-term stock performance for corrected data	
		CAR	BHAR		CAR	BHAR
CJ Korea Express	5/29/2009	-0.1513	-0.3546	DG and KV	-0.0545	-0.2734
Yuyu Pharma, Inc.	8/31/2000	1.7528	0.5037	DG	2.0489	1.1140
Namkwang Eng. and Const.	3/30/2012	9.2719	-0.9797	DG and KV	0.2983	-1.1284
Pumyang Construction	12/28/2012	69.4275	5.9651	DG and KV	-1.6690	-0.9958
Pumyang Construction	1/31/2014	-1.7378	-1.0471	DG and KV	-1.8401	-1.0179
Chinhung International	3/30/2012	-1.1531	-0.9433	DG and KV	-1.1485	-0.9427
Schnell Biopharmaceuticals	5/31/2009	21.7136	10.2138	DG and KV	0.1234	-0.6037
S&T Dynamics	3/31/2003	0.3899	-0.1936	KV	-0.0961	-0.6482
Tway Holdings	4/30/2009	2.7088	-0.0115	DG and KV	-1.9669	-1.0063
Chokwang Paint	1/31/2012	2.0122	1.8872	KV	2.2404	2.1709
Kukdong Corporation	1/31/2012	3.2059	6.8763	KV	3.4733	8.0661
Hansol Artone Paper	7/31/2009	-0.0747	-0.4550	DG and KV	-1.0392	-0.7676
Hangchang Paper	6/30/2009	5.1804	3.4290	DG and KV	-0.0743	-0.3950
Youngone Holdings	7/31/2009	1.9829	4.2510	DG and KV	1.3390	2.1937
Hyundai Paint	4/30/2014	24.8417	13.9301	DG and KV	-0.1343	-0.5100
STX Corporation	3/31/2014	2.6591	-0.7951	DG and KV	-1.3575	-1.0844
Daeyoung Packaging	12/31/2002	-0.4248	-0.9571	DG and KV	0.3187	-0.4916
Daeyoung Packaging	2/28/2003	0.0487	-0.7993	DG and KV	0.6452	-0.1241
Youngbo Chemical	1/31/2012	0.1339	-0.0911	KV	0.3541	0.0684
Iljin Display	8/31/2009	1.8067	2.5033	DG and KV	2.0125	3.7961
Maniker	10/31/2002	-0.2099	-1.0202	DG and KV	0.6074	0.1342
LG Chemical	4/30/2009	1.6142	2.8937	DG and KV	1.0840	1.5114
Artis	4/30/2012	3.8684	1.2870	DG and KV	-0.0758	-0.6103
Wooridul Huebrain	1/31/2012	1.8080	-0.5707	KV	2.0546	-0.4897
-	-	6.6948	1.8968	Mean	0.2977	0.3319
-	-	1.8074	-0.0513	Median	0.2109	-0.4907

This table presents the effect of large errors on the long-term abnormal stock returns for a sample of 24 firm-months with discrepancies greater than 20% between the DataGuide and KisValue databases. The table contrasts the long-term abnormal stock returns estimated by using erroneous monthly returns data with those estimated by using corrected monthly returns data resulting from righting 24 large errors detected in the DataGuide and/or KisValue databases. In order to measure long-term stock performance, we calculate a 36-month cumulative abnormal return (CAR) and buy-and-hold abnormal return (BHAR) using the KOSPI equally weighted market index as a return benchmark.

$$BHAR_j = \prod_{t=1}^{36} [1 + r_{jt}] - \prod_{t=1}^{36} [1 + r_{Et}] \quad [\text{Eqn 11}]$$

where

CAR_j = a 36-month CAR for a sample firm j

$BHAR_j$ = a 36-month BHAR for a sample firm j

r_{jt} = the month t actual return on a sample firm j

r_{Et} = the month t actual return on an equally weighted market index.

Table 5 shows that the mean 36-month CAR for erroneous returns data (6.6948) is more than 20 times larger than the mean 36-month CAR for corrected returns data (0.2977), while the mean 36-month BHAR for erroneous returns data (1.8968) is more than five times greater than the mean 36-month BHAR for corrected returns data (0.3319). This result suggests that large errors in the DataGuide and KisValue databases could introduce a severe upward bias in the mean CAR and BHAR, even though either an upward or a downward bias in a CAR and BHAR is found for individual stocks. In Table 5, note that these strong effects of large errors on the long-run abnormal stock returns occur even though only one of the 36 monthly stock returns used for calculating a 36-month CAR and BHAR is erroneous.

The methodological implications of errors in financial databases

As shown in Tables 4 and 5, extreme monthly returns resulting from large errors in financial databases will probably have

strong impacts both on the distributional properties of monthly stock returns and on long-term stock performances. In fact, researchers in finance are already aware of the presence of data errors in commonly used databases. Therefore, they usually try to minimise the dominant effects of extreme values by using the traditional statistical method of either trimming all values outside the bounds or by transforming them into a specified percentile of the data (winsorising).

However, using these statistical methods is not always a good practice, because outliers are not necessarily erroneous. For example, the maximum value of monthly stock returns in the DataGuide database, 7.7702, is not an erroneous, but a valid return, even though it is definitely an extreme value in comparison with the mean return, 0.0154, as shown in Table 4. This extreme return was generated by the monthly return on the common shares of Chosun Welding Co. Ltd. in March 2010, because they hit the ceiling for 17 consecutive trading days during the month (*E-Daily*, 04 March 2010). Extreme values that are not erroneous should be included in the study sample, because they represent possible states of the object studied. Moreover, deleting extreme observations is very likely to affect the output of most empirical models because of their sensitivity to outliers, and this may drastically change the results of the study (Chychyla & Kogan 2015:43–44).

Therefore, to ensure the reliability of empirical research on capital markets in Korea, it is necessary to minimise large

errors in popular financial databases such as DataGuide and KisValue that lead to extreme values or outliers. Of course, legitimate outliers that do not result from errors should be used properly as needed. If users really want high-quality databases to protect their decisions based on financial databases from being distorted by data errors, they should use cross-checking to screen for data errors in alternative financial databases on a regular basis. As a good example of how cross-checking could be used for quality control in popular databases, Bennin (1980) shows that the percentage of discrepancies, greater than 5% between CRSP and updated Compustat databases, dropped to 0.25% from 0.75% in the study by Rosenberg and Houglet (1974). That is, the Rosenberg and Houglet (1974) study comparing CRSP and Compustat databases by using cross-checking contributed to a significant improvement in the reliability of the Compustat database.

Meanwhile, companies that provide databases, such as DataGuide and KisValue in Korea, should implement their own verification systems whereby data managers screen for errors in databases by using cross-checking periodically and correct them immediately in order to maintain high-quality databases. Additionally, as most of the large errors that generate extreme values in the DataGuide and KisValue databases result from the mishandling of split events and of missing values in calculating monthly stock returns, it is most important for the database companies to educate data managers to fully understand corporate events that affect stock returns (i.e. stock splits, stock dividends, capital reductions and spin-offs) and the statistical concept of missing values.

Summary and conclusions

We examined data errors in the DataGuide and KisValue databases, which are commonly used by financial professionals in Korea, by using cross-checking. We focused mainly on comparing monthly stock returns for 729 KRX listed securities available in both the DataGuide and KisValue databases covering 15 years from January 2000 to December 2014.

We find strong evidence that there exist material discrepancies in monthly stock returns between the DataGuide and KisValue databases, most of which are attributable to errors in handling stock split events and missing returns. Specifically, out of 109 260 comparisons, we find 2563 (2.35%) to be erroneous, including 58 (0.05%) that differ by more than 5% and 24 (0.02%) that differ by more than 20%. In addition, the mishandling of split events (i.e. stock splits, reverse stock splits, rights offering and spin-offs) causes serious data errors in proportion to the split ratios which range from 2 to 72. The results also show a DataGuide error rate of 0.03% and a KisValue error rate of 0.41%, indicating that DataGuide is a more reliable database than KisValue for monthly stock returns. Further, the results also show that all the extreme returns, which from large errors in the two databases, can be significant enough to sharply change the properties of monthly stock return distributions and to over- or

underestimate long-run abnormal stock returns. In particular, the mean 36-month CARs estimated, which used erroneous monthly returns data, are 20 times larger than those estimated using corrected returns data for the 24 firm-month sample in which DataGuide and/or KisValue made serious errors.

Finance researchers in Korea already know that there are data errors in popular financial databases such as DataGuide and/or KisValue. They assume that outliers might be erroneous and could have a significant effect on empirical analysis. Therefore, in order to minimise the effect of outliers, they discard them or transform them using the winsorising method. However, using these statistical methods is not always a good practice, because all outliers are not necessarily erroneous. In this regard, the users of financial databases must regularly examine data errors even in highly reputed databases and ask database companies to correct the errors in order to ensure reliable financial databases in Korea. On the contrary, database companies should develop a more sophisticated algorithm to take into account the effect of split events in calculating monthly stock returns. Further, they also need to introduce a series of special missing-return codes, specifying the reason for missing returns as in the CRSP database. Finally, not only users, but also database companies have to keep in mind that 'the presence of erroneous data can destroy a research effort and seriously damage the management decisions based upon research', as stated by Rosenberg and Houglet (1974).

Acknowledgements

The authors would like to thank an anonymous referee for helpful comments. The authors also would like to thank the Pukyong National University for funding this study.

This work was supported by the Pukyong National University Research Abroad Fund in 2015 (C-D-2015-0511).

Competing interests

The authors declare that they have no financial or personal relationships which may have inappropriately influenced them in writing this article.

Authors' contributions

H.-C.J. and H.-J.N. conceived and designed the research. H.-J.N. undertook the primary research and H.-C.J. was in charge of analysing the data and contributed to organising all sections and critical revision. Both authors read and approved the final manuscript.

References

- Baik, B., Kang, H. & Kim, Y.J., 2013, 'Volatility arbitrage around earnings announcements: Evidence from the Korean equity linked warrants market', *Pacific-Basin Finance Journal* 23, 109–130. <https://doi.org/10.1016/j.pacfin.2013.01.001>
- Baik, B., Kim, Y.J., Kim, J. & Lee, S.J., 2015, 'Usefulness of earnings in credit markets: Korean evidence', *Pacific-Basin Finance Journal* 33, 93–113. <https://doi.org/10.1016/j.pacfin.2015.01.009>
- Beedles, L. & Simkowitz, M., 1978, 'A note on skewness and data errors', *The Journal of Finance* 33(1), 288–292. <https://doi.org/10.1111/j.1540-6261.1978.tb03405.x>

- Bennin, R., 1980, 'Error rates in CRSP and Compustat: A second look', *The Journal of Finance* 35(5), 1267–1271. <https://doi.org/10.1111/j.1540-6261.1980.tb02210.x>
- Center for Research in Security Prices (CRSP), 2012, *Data descriptions guide: US Stock & US index databases*, Chicago Booth, Chicago, IL.
- Chang, J. & Shin, H.H., 2007, 'Family ownership and performance in Korean conglomerates', *Pacific-Basin Finance Journal* 15, 329–352. <https://doi.org/10.1016/j.pacfin.2006.07.004>
- Choi, I.G., Sohn, P. & Seo, J.Y., 2015, 'The relationship between labour unions' bargaining power and firms' operating flexibility: New evidence from emerging markets', *South African Journal of Business Management* 46(4), 65–75. <https://doi.org/10.4102/sajbm.v46i4.110>
- Chychyla, R. & Kogan, A., 2015, 'Using XBRL to conduct a large-scale study of discrepancies between the accounting numbers in Compustat and SEC 10-K filings', *Journal of Information Systems* 29(1), 37–72. <https://doi.org/10.2308/isys-50922>
- Dataguide.co.kr, FnGuide Inc., Seoul, Korea, updated 30 June 2015, viewed 30 June 2015, from <http://www.dataguide.co.kr/>
- Jung, H.C., 2007, 'The power and misspecification of the models measuring long-run security price performance: The case of Korea stock market', *Asia-Pacific Journal of Financial Studies* 36(2), 237–280.
- Jung, H.C., 2010, 'Valuation effects of private and public target mergers in Korea', *Asia-Pacific Journal of Financial Studies* 39(6), 752–776. <https://doi.org/10.1111/j.2041-6156.2010.01027.x>
- Kho, B.C. & Park, R.S., 2000, 'An empirical analysis on the abnormal performance of security-issuing firms in Korea', *Korean Journal of Financial Studies* 27, 439–476.
- Kim, S.H., Kim, D. & Shin, H.S., 2012, 'Evaluating asset pricing models in the Korean stock market', *Pacific-Basin Finance Journal* 20(2), 198–227. <https://doi.org/10.1016/j.pacfin.2011.09.001>
- Kim, S.Y., Lee, K.R. & Shin, H.H., 2017, 'The enhanced disclosure of executive compensation in Korea', *Pacific-Basin Finance Journal* 43, 72–83. <https://doi.org/10.1016/j.pacfin.2017.02.005>
- Kinney, M. & Swanson, E., 1992, 'The accuracy and adequacy of tax data in Compustat', *Journal of the American Taxation Association* 15(1), 121–135.
- Kisvalue.com, NICE Information Service, Seoul, Korea, updated 03 July 2015, viewed 03 July 2015, from <http://www.kisvalue.com/>
- Koo, B.Y., 1998, 'A test of efficiency of proxy market portfolios by generalized method of moments: Evidence from the Korean stock market', *Korean Journal of Financial Management* 15(1), 1–30.
- Lee, D. & Cho, J., 2014, 'Stock price reactions to news and the momentum effect in the Korean stock market', *Asia-Pacific Journal of Financial Studies* 43(4), 556–588. <https://doi.org/10.1111/ajfs.12058>
- Nam, H.J., 2016, 'A study on the quality of financial information provided by DataGuide and TS2000', *Journal of the Korean Data Analysis Society* 18(4), 2053–2065.
- Oh, M. & Lee, E., 2007, 'A comparative study on accounting numbers of financial databases', *Korean Journal of Business Administration* 43(4), 2955–2978.
- Rosenberg, B. & Hougllet, M., 1974, 'Error rates in CRSP and Compustat data bases and their implications', *The Journal of Finance* 29(4), 1303–1310. <https://doi.org/10.1111/j.1540-6261.1974.tb03107.x>
- San Miguel, J., 1977, 'The reliability of R&D data in Compustat and 10-K reports', *The Accounting Review* 52(3), 638–641.
- Tallapally, P., Luehlfling, M. & Motha, M., 2011, 'The partnership of EDGAR online and XBRL-should Compustat care?', *Review of Business Information System* 15(4), 39–46. <https://doi.org/10.19030/rbis.v15i4.6011>
- Winkler, J., Kuklinski, C.P.J.-W. & Moser, R., 2015, 'Decision making in emerging markets: The Delphi approach's contribution to coping with uncertainty and equivocality', *Journal of Business Research* 68(5), 1118–1126. <https://doi.org/10.1016/j.jbusres.2014.11.001>
- Yang, D., Vasarhelyi, M. & Liu, C., 2003, 'A note on the using of accounting databases', *Industrial Management and Data Systems* 103(3), 204–210. <https://doi.org/10.1108/02635570310465689>