Marais, Alastair; Vermaak, Claire; Shewell, Patricia

Article

Predicting financial statement manipulation in South Africa: A comparison of the Beneish and Dechow models

Cogent Economics & Finance

Provided in Cooperation with:
Taylor & Francis Group

Suggested Citation: Marais, Alastair; Vermaak, Claire; Shewell, Patricia (2023) : Predicting financial statement manipulation in South Africa: A comparison of the Beneish and Dechow models, Cogent Economics & Finance, ISSN 2332-2039, Taylor & Francis, Abingdon, Vol. 11, Iss. 1, pp. 1-33, https://doi.org/10.1080/23322039.2023.2190215

This Version is available at:
https://hdl.handle.net/10419/304022

Cogent Economics & Finance

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/oaef20

Predicting financial statement manipulation in South Africa: A comparison of the Beneish and Dechow models

Alastair Marais, Claire Vermaak & Patricia Shewell

To cite this article: Alastair Marais, Claire Vermaak & Patricia Shewell (2023) Predicting financial statement manipulation in South Africa: A comparison of the Beneish and Dechow models, Cogent Economics & Finance, 11:1, 2190215, DOI: 10.1080/23322039.2023.2190215

To link to this article: https://doi.org/10.1080/23322039.2023.2190215

Published online: 15 Mar 2023.

Marais et al., *Cogent Economics & Finance* (2023), 11: 2190215
https://doi.org/10.1080/23322039.2023.2190215

cogent economics & finance

FINANCIAL ECONOMICS | RESEARCH ARTICLE

Predicting financial statement manipulation in South Africa: A comparison of the Beneish and Dechow models

Alastair Marais[1]*, Claire Vermaak[1] and Patricia Shewell[1]

Received: 17 November 2022
Accepted: 08 March 2023

*Corresponding author: Alastair Marais, School of Accounting, Economics and Finance; University of KwaZulu-Natal, Private Bag X54001, Durban, 4000, KwaZulu-Natal, South Africa
E-mail: maraisa@ukzn.ac.za

Reviewing editor:
David McMillan, Accounting and Finance, University of Stirling, Stirling, United Kingdom

Additional information is available at the end of the article

Abstract: Recently, South Africa has suffered from several large financial statement frauds. To assist stakeholders in identifying fraud, this study investigated the ability of the Beneish M-score and the Dechow *et al*. F-score to identify fraud in South Africa. The study also explored similarities in earnings management characteristics between false positives and fraudulent companies. Finally, the study re-estimated the models' coefficients based on current South African data to determine if this improved their predictive capabilities. The study used a sample of 23 manipulated and 2 320 non-manipulated observations from 2006 to 2018 and found that both scores showed low sensitivity and precision. The false positives share similar, or higher, earnings management characteristics to the manipulators. Re-estimating the coefficients reduced the M-scores' sensitivity by, on average, 6.52% but improved precision by, on average, 4.21%. Conversely, re-estimation increased the F-scores' sensitivity by, on average, 58.70% but increased the type II error by, on average, 48.09%. These findings suggested that either the M- and F-scores are unsuitable in the South African context or that regulators have failed to identify manipulators adequately. Therefore, investors and other stakeholders should use caution when applying these models in South Africa.

Subjects: Auditing; Business Ethics; Corporate Governance

Keywords: Beneish M-score; Dechow *et al*. F-score; earnings management; fraud detection; fraudulent financial reporting; South Africa

JEL Classifications: G11; G32; G38; M41; M42

1. Introduction

Globally, financial statement fraud accounts for ten per cent of occupational frauds (Association of Certified Fraud Examiners, 2020). While this is the least common of the three major fraud categories (asset misappropriation, corruption and financial statement fraud), it is the costliest, resulting in a median loss of United States (US) $954 000 in 2020 (Association of Certified Fraud Examiners, 2020). Concerningly, an increase in financial statement fraud is anticipated in the post-COVID-19 pandemic period (Association of Certified Fraud Examiners, 2021). Financial statement fraud undermines the quality of financial data utilised to make economic decisions. Poor economic decisions lead to financial loss for the stakeholders and may have negative consequences for an economy due to an inefficient allocation of resources (Pududu & De Villiers, 2016).

South Africa (SA) is no stranger to financial statement fraud, with the Steinhoff and Tongaat-Hulett scandals being two of the largest frauds in recent years. The Steinhoff scandal broke in December 2017 with the resignation of then-CEO Markus Jooste and the commencement of an investigation into accounting irregularities, including overstating revenue and hiding losses in off-balance-sheet companies (Hlobo et al., 2022; Rossouw & Styan, 2019). These revelations resulted in the share price declining from ZAR45.65 at the start of trading on 6 December 2017 to ZAR17.61 by the close of the day (Van Der Linde, 2022). By the close of trading on 8 December 2017, the share price had declined to ZAR6.00 and continued to descend (Rossouw & Styan, 2019). A few months later, in 2018, fraud at Tongaat-Hulett was revealed. The company's financial results had been overstated by approximately ZAR4.5 billion through the overstatement of revenue and assets and the understatement of expenses (Hlobo et al., 2022; Muzata & Marozva, 2022).

In addition to the frauds mentioned above, the PricewaterhouseCoopers (2020) Global Economic Crime and Fraud survey reported that SA had the third-highest occurrence of economic crime in the world, after India and China. The survey revealed that 60% of SA companies had been affected by fraud or economic crime between 2009 and 2020, compared to 47% of companies globally. The survey indicated that the percentage of companies experiencing accounting and financial statement fraud in SA had increased from 22% in 2018 to 34% in 2020.

Notwithstanding the prevalence of fraud in SA and the related economic costs, companies' responses to fraud prevention and detection have been ineffective. In Sub-Saharan Africa (a region that includes SA), an external audit is the most common anti-fraud control, despite only being responsible for the initial detection of 4% of frauds (Association of Certified Fraud Examiners, 2020). Concerningly, several financial statement frauds are committed with the auditor's knowledge (Mongwe & Malan, 2020). Only 58% of companies in SA reported having performed an investigation of their most severe fraud, and 59% of such frauds were never reported to the board of directors, 66% were not reported to the appropriate regulator, and 72% were never disclosed to the auditors (PricewaterhouseCoopers, 2020).

Given the country's high levels of fraud, SA provides a unique environment to study the detection of financial statement fraud. The country has the third-largest economy in Africa and is an emerging economy (World Bank, 2020), and is characterised by a small stock exchange, an insider economy, concentrated ownership and weak legal enforcement. These factors increase the risk of fraud (Pududu & De Villiers, 2016). Although investors in a high-risk country should be able to better detect manipulated financial statements, SA investors struggle to do so (Rabin, 2016). Despite these negative characteristics, SA has, until 2017, consistently ranked highly in the World Economic Forum's (2017) Global Competitiveness Report in terms of strong investor rights, the strength of auditing and financial reporting standards, protection of minority shareholders, efficiency of corporate boards and firm ethical behaviour. Following the revelations around the Steinhoff scandal, SA's rankings in the Global Competitiveness Reports declined markedly post-2017.

Academic research on fraud detection in SA is limited. An early study by Koornhof and Du Plessis (2000) considered red flags as an early warning system to identify potential fraud. A series of articles used machine learning models to identify qualified audit opinions (see Moepya et al., 2016; Moepya, Akhoury, et al., 2014; Moepya, Nelwamondo, et al., 2014). Finally, Rabin (2016) used earnings discontinuities to identify companies engaging in earnings management, a precursor to financial statement fraud (Mishra & Malhotra, 2016).

Given SA's prevalence of fraud, the inability of investors to detect misrepresented financial statements, and the limited academic literature on fraud detection, the purpose of this study was three-fold. The first objective used a sample of 23 manipulated observations and 2 320 non-manipulated observations from 2006 to 2018 to determine the usefulness of two popular financial statement fraud detection models (namely the Beneish (1999) M-score and the Dechow et al.

(2011) F-score) to detect cases of fraud in SA correctly. Several recent academic studies in African countries have used these models as proxies for financial statement fraud risk (see, for example, Mavengere (2015), Nyakarimi (2022), Nyakarimi et al. (2020), Okiro and Otiso (2021), Onyebuchi and Nkem (2021)). However, few studies have thoroughly tested the models' prediction abilities in contexts outside the United States and, specifically, in Africa. Consequently, Rad et al. (2021) call on researchers to test the accuracy of fraud detection models to determine their effectiveness in the context they are applied. This is particularly relevant given that both models were developed in the US using pre-2005 data. As South Africa is considered an emerging economy and uses International Financial Reporting Standards (IFRS), it provides a very different context to the US, a developed country that uses US GAAP.

The second objective was to investigate the nature of the false positives produced by these models to determine whether they have similar earnings management characteristics to fraudulent companies. Concerningly, prediction models tend to generate many false positives (Beneish & Vorst, 2021; Walker, 2020). Consequently, Dechow et al. (2011) call for further research into the characteristics of false positives, but research in this area is limited. The final objective was to re-estimate the coefficients of the two models based on SA data to increase the models' predictive ability in SA. This addresses initial concerns about the differences in the US (where the models were developed) and SA contexts, as well as the later period under consideration.

This study contributes to the existing body of knowledge by showing that the M- and F-scores perform poorly in correctly identifying manipulating companies in SA. African studies have incorrectly relied on the earlier good performance of these models in non-African contexts (such as the US, Europe and Asia) without testing the validity of the models in an African context. Of further concern is that recent studies in the US and China have shown declined performance of these prediction models (see, for example, Beneish and Vorst (2021) and Lu and Zhao (2021)), highlighting the need to test these models thoroughly in different contexts before relying on them. The study further contributes to understanding the nature of false positives generated by the models. In this study, false positives were shown to have similar or higher levels of accruals-based earnings management compared to the manipulator sample, highlighting that the models may not be picking up fraud but rather aggressive accounting practices. Finally, the study further provides evidence that re-estimation may not improve the models' performance. Re-estimation of the M-score coefficients using publicly-available South African data reduces the models' ability to identify manipulators correctly by, on average, 6.52%. Conversely, re-estimating the coefficients for the F-score improves the scores' ability to classify manipulators correctly by, on average, 58.70%, but the number of false positives is substantially increased.

The remainder of the article is organised as follows: section 2 presents the literature review and hypothesis development, section 3 details the methodology applied in the study, the results are presented and discussed in sections 4 and 5 and, finally, section 6 concludes.

2. Literature review and hypothesis development

2.1. Defining financial statement fraud

Financial statement fraud is defined as an intentional misstatement of financial statements to gain some benefit (Association of Certified Fraud Examiners, 2020). It is essential to distinguish between financial statement fraud and earnings management. While both relate to intentional misstatement for economic gain, financial statement fraud occurs outside acceptable accounting standards, while earnings management occurs within such standards (Albizri et al., 2019).

2.2. Financial statement fraud detection models

Financial statement fraud detection models incorporate financial ratios and other elements, such as textual analysis, which contain proxies for the fraud risk factors identified in the theoretical literature. Models have been developed using various methods, including simple financial ratios,

statistical methods (such as logit and probit models) and advanced machine learning methods (such as artificial neural networks and support vector machines). While neural networks are the most widely used method in the academic literature, they are complex, lack transparency and are less interpretable (Mongwe & Malan, 2020). As a result, they are not suitable for widespread use in emerging markets such as SA. In addition, these advanced methods do not necessarily deliver superior predictive power than the F-score or a simple screen of sales growth (Walker, 2020). Mongwe and Malan (2020) claim that there is no overall best method, with performance often based on the data set used.

For these reasons, this study used the M- and F-scores. Both models are widely used in the literature and require only information directly obtainable from the company financial statements to estimate. They can thus serve as suitable screening tools (Skousen & Twedt, 2009), particularly in emerging economies where there is increased information asymmetry and a lack of comprehensive databases compared to advanced economies. In addition, the F-score is considered the standard in financial statement fraud prediction (Walker, 2020).

2.2.1. Beneish (1999) M-score

The M-Score was developed by Beneish (1999) using probit estimation with data from 1982 until 1992. US Security Exchange Commission (SEC) enforcement actions and news reports were used to identify 74 non-financial US companies that manipulated their earnings matched to 2 332 non-manipulators by industry and year. The financial statement elements used to predict manipulation were based on signals identified in the academic and practitioner literature. The unweighted model, as estimated by Beneish (1999), is as follows:

$$M = -4.840 + 0.920DSRI + 0.528GMI + 0.404AQI + 0.892SGI + 0.115DEPI - 0.172SGAI + 4.679TATA - 0.327LVGI \quad (1)$$

Where *DSRI* refers to the days' sales in receivables index, *GMI* refers to the gross margin index, *AQI* denotes the asset quality index, *SGI* denotes the sales growth index, *DEPI* is the depreciation index, *SGAI* is the sales, general and administrative expenses index, *TATA* refers to the total accruals to total assets, and finally, *LVGI* is the leverage index (Beneish, 1999). The detailed variable calculations are presented in Appendix 1. Beneish (1999) then determined three cut-off points that minimised the expected cost of misclassification (ECM) at different relevant costs of type I and II errors.[1] These cut-off points were −1.49, −1.78 and −1.89, representing relative costs of 10:1, 20:1 and 40:1, respectively, where a score greater than the cut-off indicates that the company is classified as a manipulator.

Not all the variables in the M-score are equally important (Paolone & Magazzino, 2015). As a result, a simplified five-variable model was also developed in the literature as follows (Nyakarimi, 2022):

$$M = -6.065 + 0.823DSRI + 0.906GMI + 0.593AQI + 0.717SGI + 0107DEPI \quad (2)$$

Where the variables maintain their meaning from the original model. However, as the full M-score has not yet been thoroughly tested in the South African context, and in line with the majority of academic literature (see, for example, Aghghaleh et al. (2016), M. D. Beneish et al. (2013), D. Beneish and Vorst (2021), Cecchini et al. (2010), Jones et al. (2008), Kamal et al. (2016), Price et al. (2011) and Nurul Herawati (2015), this study uses the original M-score, inclusive of all eight variables.

2.2.2. Dechow et al. (2011) F-score

Dechow et al. (2011) also recognised the usefulness of financial information beyond accruals to detect financial statement fraud. Unlike prior models, however, they aimed to allow a user to

calculate the F-score for an individual company and simplify the assessment of whether it was misstating its financial statements. To achieve this, they did not include any indices as their variables or perform any form of matching between manipulating and non-manipulating firms. Using a total of 2 190 accounting violations identified by the US SEC from May 1982 to June 2005, they developed three models using logistic regression to detect manipulation. Model 1 contained financial statement variables only as follows:

$$\begin{aligned} \textit{Predicted value} = & -7.893 + 0.790RSST + 2.518\Delta REC + 1.191\Delta INV + 1.979SASS \\ & + 0.171\Delta CSALES - 0.932\Delta ROA + 1.029AISS \end{aligned} \tag{3}$$

Where *RSST* represents accruals as measured by Richardson et al. (2005),[2] *ΔREC* is the change in receivables, *ΔINV* is the change in inventory, *SASS* is the percentage of soft assets,[3] *ΔCSALES* represents the percentage change in cash sales, *ΔROA* is the change in return on assets, and finally, *AISS* represents whether the company issued securities during the period (Dechow et al., 2011). Appendix 2 presents full details of the variable calculations.

Model 2 introduced off-balance sheet and non-financial variables as follows:

$$\begin{aligned} \textit{Predicted value} = & -8.252 + 0.665RSST + 2.457\Delta REC + 1.393\Delta INV + 2.011SASS \\ & + 0.159\Delta CSALES - 1.029\Delta ROA + 0.983AISS - 0.150\Delta EMP + 0.419LEASE \end{aligned} \tag{4}$$

Where *ΔEMP* and *LEASE* represent the abnormal change in employees and the existence of operating leases, respectively (Dechow et al., 2011).

Finally, Model 3 added two market-based variables as follows:

$$\begin{aligned} \textit{Predicted value} = & -7.966 + 0.909RSST + 1.731\Delta REC + 1.447\Delta INV + 2.265SASS \\ & + 0.160\Delta CSALES - 1.455\Delta ROA + 0.651AISS - 0.121\Delta EMP + 0.345LEASE \\ & + 0.082RET_t + 0.098RET_{t-1} \end{aligned} \tag{5}$$

Where RET_t and RET_{t-1} represent the market-adjusted share returns and the lagged market-adjusted share returns, respectively (Dechow et al., 2011).

The first model offers two advantages. First, it contains most of the predictive power. Second, it is the least restrictive model, as the required information may be accessed from financial statements (Price et al., 2011; Skousen & Twedt, 2009). This second benefit is particularly relevant for emerging economies. Thus, given the importance of this second benefit for the current study's context, as well as in line with the majority of the prior literature (see, for example, Aghghaleh et al. (2016), Chakrabarty et al. (2022), Price et al. (2011) and Walker (2020)), Model 1 of the F-score is used in this study.

Following the calculation of the predicted value, it is then converted to a probability as follows:

$$\textit{Probability} = \frac{e^{\textit{Predicted value}}}{1 + e^{\textit{Predicted value}}} \tag{6}$$

Finally, the F-score is calculated by dividing the probability by the "unconditional expectation of misstatement" (UEM). The UEM is the proportion of misstated firms to total firms (Dechow et al., 2011:60). Companies that obtained an F-score above one are considered an above-normal risk, whilst companies scoring above 2.45 have a high risk of manipulation (Dechow et al., 2011).

2.2.3. Comparative performance literature

Numerous studies have investigated the ability of the M- and F-scores to detect financial statement fraud. In his original study, Beneish (1999) determined that the M-score could correctly detect 76% of manipulating firms and 82.5% of non-manipulating companies in the estimation sample. The model only identified 56.1% of manipulators in the holdout sample, although the correct classification of non-manipulating companies rose to 90.9%. Several later studies also found positive results for the model. In the US, using a maximum of 142 manipulated and 72 815 non-manipulated observations from 1988 to 2001, Jones et al. (2008) found that the model was significantly positively associated with both the occurrence of fraud and the magnitude of the fraud. Using a later sample of 43 534 US observations over the period 1993 until 2010, Beneish et al. (2013) showed that the M-score could identify 71% of manipulators. In Asia, Tarjo and Herawati (2015) used a matched sample (based on assets and industry) of 35 manipulators and 35 non-manipulators from 2001 to 2014. They found that 77.1% of the manipulators and 80% of the non-manipulators were correctly classified. In Malaysia, Kamal et al. (2016) tested the M-score's ability to identify 17 manipulated companies from 1993 to 2014. They reported an 82% accuracy when using a −2.22 cut-off, a 76% accuracy for a −1.89 cut-off and a 71% accuracy for the −1.78 cut-off.

Regarding the F-score, in their original study, Dechow et al. (2011) identified that Model 1 correctly classified 68.6% of manipulating companies and 63.7% of non-manipulators in the estimation sample and 73.8% of manipulating companies and 61.7% of non-manipulating companies in the holdout sample. A subsequent study in the US from 1991 until 2008 by Chakrabarty et al. (2022) used a sample of 853 manipulators and 119 967 non-manipulators. They found that the F-score correctly identified 68.5% of manipulators and 57.5% of non-manipulators.

Based on the above results and the detective power of the M- and F-scores, recent African literature has relied on these models as proxies for fraud (see, for example, Mavengere (2015), Nyakarimi (2022), Nyakarimi et al. (2020), Okiro and Otiso (2021), Onyebuchi and Nkem (2021)). However, these studies ignore that these models have not been tested in the African context, where they may not be applicable due to the different context from the US and the later period (Lu & Zhao, 2021). Further, more recent studies have found that the models, particularly the M-score, are less able to predict manipulation in recent times correctly. For example, Beneish and Vorst (2021) used a sample of 768 manipulated observations and 136 144 non-manipulated observations from 1979 to 2016 in the US. They found that the M-score only identified 23.18% of manipulators. Likewise, Lu and Zhao (2021) randomly selected 40% of a sample of 190 manipulators and 9 693 non-manipulators for Chinese listed firms. They found that the M-score could only detect 29.63% of the fraud sample.

Thus, given the mixed findings and the seeming decline in the models' performance, there is a need to test whether the M- and F-scores are relevant in the SA context before being able to rely on the models as proxies for fraud risk. Consequently, the following hypothesis is drawn:

H1: The M- and F-scores can detect financial statement fraud in SA.

Several studies have compared the performance of the M- and F-scores on a homogenous sample. These studies have demonstrated that, while both models can correctly identify manipulating companies, the F-score is a more robust model with greater predictive accuracy. Cecchini et al. (2010) used US data from 1991 to 2003. Using 149 fraudulent observations matched to 3 389 non-fraudulent observations (based on industry and year), they found that the M-score correctly classified 54.2% of fraudulent and 45.5% of non-fraudulent observations. Using 57 fraudulent and 1 244 non-fraudulent observations,[4] the F-score outperformed the M-score by correctly identifying 70.0% of fraudulent and 84.9% of non-fraudulent observations. Price et al. (2011) also studied US companies. They used a total sample of 57 185 observations from 1994 until

2008, including 866 SEC enforcement actions, 542 accounting irregularities and 948 lawsuits. Their results found that the F-score outperformed the M-score. In a Malaysian context, Aghghaleh et al. (2016) used a one-for-one matched sample (based on industry and year) of 82 fraudulent observations from 2001 to 2014. They found that the F-score identified a higher proportion of fraudulent observations than the M-score (73.17% compared to 69.51%) with a lower type II error (26.83% compared to 30.49%).

Based on these studies, the F-score seems to have greater detecting power than the M-score. Therefore, the following hypothesis is drawn:

H2: The F-score outperforms the M-score in detecting financial statement fraud in SA.

2.3. Earnings management characteristics of false positives

A fundamental problem with financial statement fraud detection models is the high occurrence of type II errors (false positives) generated (Beneish & Vorst, 2021). This problem is particularly prevalent when detecting a rare event such as financial statement fraud (Walker, 2020). Given the inherent unobservability of financial statement fraud and the resource constraints regulators face when investigating such fraud, an avenue for further research is identifying characteristics of the false positives (Dechow et al., 2011).

Multiple studies revealed that companies that commit fraud have previously engaged more aggressively in earnings management (Dechow et al., 1996; Marinakis, 2011; Perols & Lougee, 2011). As extensive earnings management eventually reverses or reduces manipulation flexibility, managers may resort to fraud to maintain appearances (Perols & Lougee, 2011). For this reason, earnings management is considered a precursor to accounting fraud (Mishra & Malhotra, 2016). Therefore, it is expected that companies identified as false positives by the M- and F-scores would display earnings management characteristics more in accordance with the manipulator sample. Hypothesis three is thus:

H3: The false positive samples generated by the M- and F-scores display earnings management characteristics consistent with the manipulator sample.

2.4. M-score, F-score and model drift

The M- and F-scores were developed in the US using pre-2005 data. These models are static; the world, however, changes. Thus, using these models on more recent data in a different country may reveal model deterioration (Lu & Zhao, 2021). This is due to either concept drift (where the output characteristics change) or data drift (where the input characteristics change) (Ackerman et al., 2019; Webb et al., 2016).

Several studies have updated the M- and F-scores in different ways. First, some studies (such as Cecchini et al. (2010) and Marinakis (2011)) re-estimated the coefficients using US data from 1991 to 2003 and UK data from 1994 to 2007, respectively. Next, other studies (such as Hung et al. (2017) and Putra and Dinarjito (2021), who studied 614 Vietnamese observations from 2014 to 2016 and 81 Indonesian companies from 2012 to 2018, respectively) first identified variables within the scores which could differentiate between manipulators and non-manipulators. Variables that were unable to differentiate were omitted from the models before re-estimating the coefficients. The last group of studies (such as Chakrabarty et al. (2022), Hung et al. (2017), Lu and Zhao (2021) and Marinakis (2011)) added additional variables in an attempt to improve the models. While most of these studies do not report a direct comparison between the predictive ability of the original and revised models, Chakrabarty et al. (2022) noted that, for the estimation and holdout sample, the model's ability to correctly detect manipulators increases by 3.6% and 3% respectively after the inclusion of additional variables and re-estimation[5].

The following research hypothesis is, therefore, developed:

H4: Updating the coefficients of the M- and F-scores will increase the ability of the two models to identify manipulators and decrease misclassification errors in SA.

3. Methodology

3.1. Population, sample and data collection

The population for this study is all 330 non-financial companies listed on the main board of the Johannesburg Stock Exchange (JSE) from 2006 until 2018. Financial companies are excluded, because the M-score was developed on non-financial firms (Kukreja et al., 2020) and financial firms have different regulatory and other requirements which may influence the outcome of the calculations (Orazalin & Akhmetzhanov, 2018). The 2006 year represents the first available enforcement action by the Financial Sector Conduct Authority (FSCA). Ending the sample in 2018 allows regulators sufficient time to investigate suspected irregularities. Walker (2020) notes that the mean and median time between the fraud and the SEC issuing an enforcement action is four years in the US. Based on 1 243 SEC enforcement actions, Karpoff et al. (2017) found the median period from the violation until the first enforcement action was 2.41 years. Finally, Bao et al. (2020) allowed for a two-year gap. In SA, studies have used other measures, such as qualified audit opinions (Moepya, 2017), small losses (Pududu & De Villiers, 2016) and earnings distribution discontinuities (Rabin, 2016), rather than enforcement actions to proxy for financial statement manipulation. Consequently, there is a lack of data on how long the Financial Reporting Investigation Panel (FRIP) and FSCA take to issue an enforcement action or equivalent. Thus, this study allowed for a three-and-a-half-year gap for regulators to identify violations (2019 until mid-2022).

Based on the above, an unbalanced panel of 330 firms across 13 years, representing 2 775 firm-year observations, was arrived at. The financial data were collected from the Standard and Poor's Capital IQ and Bloomberg databases. The "as originally reported" data was used to avoid the risk of abnormalities being removed when the data was restated.

In arriving at the final sample, 26 firm-year observations in which the company listed after year-end but before the release of the annual report were removed. Further, 52 firm-years in which a company's year-end changed were removed together with the year immediately after the change in year-end (for a total of 104 firm-years). This was due to the length of the periods not being comparable. Next, five firm-years were removed because the financial statements were reported in a currency experiencing hyperinflation.

A total of 272 observations with missing data that prevented the calculation of either the M- or F-scores were removed from the sample. Only using observations for which both models can be calculated increases the power of the statistical tests (Price et al., 2011). Mongwe and Malan (2020) also found that 94% of studies surveyed on fraud detection either do not deal with missing data or simply delete the affected observation. While this approach results in data loss, it avoids imputing data that may not exist (Mongwe & Malan, 2020).

Finally, 25 companies with only one firm-year observation were removed from the sample. This process resulted in a final unbalanced panel of 274 companies representing 2 343 firm-years, summarised in Table 1 below.

3.2. Identifying earnings manipulators

In SA, a complete list of firms that have manipulated their earnings is not readily available. Further, unlike advanced economies, the oversight bodies are not considered sophisticated and do not

Table 1. Sample size determination

	No. of companies[1]	No. of firm-year observations
Population	330	2 775
Listed after year-end but before the release of AFS	(1)	(26)
Change of year-end	(3)	(104)
Other anomalous situations	(0)	(5)
Missing data for M- or F-score	(27)	(272)
Companies with only one observation	(25)	(25)
Sample size	274	2 343

Note: [1]The removal of firm-year observations exceeds the removal of companies as, for some companies, not all observations were removed.

(Source: Researchers' own construction)

examine IFRS compliance on a sufficiently regular basis (Rabin, 2016). As such, a list of instances when companies engaged in manipulation was compiled as described below.

Investigations by regulators (such as the SEC in the US) are the most common proxy for financial statement fraud (Mongwe & Malan, 2020). SA has two regulatory bodies that monitor listed company financial statements: the FSCA (formerly the Financial Services Board) and the FRIP (formerly the GAAP Monitoring Panel). The FSCA is responsible for regulation and supervision within the SA financial markets and addresses issues around market abuses, including prohibited trading practices, insider trading and false and misleading reporting. As this study focused on financial statement fraud, only those enforcement actions relating to section 76 of the Securities Services Act no. 36 of 2004 (pre-2013) and section 81 of the Financial Markets Act no. 19 of 2012 (post-2013) were used. FSCA enforcement actions were obtained from the FSCA website.

The JSE tasks the FRIP to investigate instances of non-compliance with IFRS. Unlike the FSCA, the FRIP does not publish a list of investigations and their outcomes. However, following the investigation, the JSE may instruct companies guilty of non-compliance to publish or reissue any necessary information and make a public announcement via the Securities Exchange News Service (SENS) (Watson & Rossouw, 2012). Following Watson and Rossouw (2012), the IRESS database was searched to identify SENS announcements which included the words "GAAP Monitoring Panel", "GMP", "Financial Reporting Investigation Panel" and "FRIP". Each FSCA enforcement action and FRIP restatement identified was then examined to determine whether it involved an IFRS violation and the year(s) to which that violation relates.

Finally, similar to Moepya (2017), companies that had a qualified audit opinion during the period were included in the manipulator sample. However, not all qualifications relate to fraud (Jones et al., 2008). Thus, unlike Moepya (2017), this study excluded the emphasis of matter opinions and qualifications that did not relate to IFRS violations (i.e. going concern issues). Thus, only qualifications related to IFRS violations and disclaimers of opinion, where the auditor cannot draw an opinion, formed part of the manipulator sample.

Thus, only companies found guilty of fraud or a violation by the FSCA or FRIP, or having received a qualified audit report due to fraud or violation, were included in the manipulator sample. All other non-financial companies listed on the JSE between 1 January 2006 and 31 December 2018 formed part of the non-manipulating sample (i.e. these companies had not been found guilty of fraud or a violation, nor had they received a relevant qualified audit opinion). Table 2 discloses a sample of 23 manipulated firm-year observations (9 unique companies) representing 0.98% of the total observations. This provides a smaller absolute number of manipulated observations

Table 2. Summary of classification between manipulated and non-manipulated observations

	Current study		Beneish (1999)		Dechow et al. (2011)	
	No. of observations	**% of observations**	**No. of observations**	**% of observations**	**No. of observations**	**% of observations**
Manipulated observations	**23**	**0.98**	**74**	**3.08**	**494**	**0.37**
FSCA enforcement actions	14	0.60				
FRIP restatements	5	0.21				
Applicable qualified audit opinions	4	0.17				
Non-manipulated observations	**2 320**	**99.02**	**2 332**	**96.92**	**132 967**	**99.63**
Total observations	**2 343**	**100.00**	**2 406**	**100.00**	**133 461**	**100.00**

Note: (Source: Researchers' own construction, as well as data obtained from Beneish (1999) and Dechow et al. (2011))

cogent economics & finance

compared to the original studies. Proportionally, however, this sample does compare favourably to the original studies, particularly that of Dechow et al. (2011).

3.3. Calculation of the M- and F-scores

The M- and F-scores were estimated as described in Equations (1) and (3) above. As justified under sections 2.2.1 and 2.2.2, the original eight variable M-score and model 1 for the F-score were used. Following Beneish (1999) and Dechow et al. (2011), all variables used in calculating the M- and F-scores were winsorized at the first and ninety-ninth percentiles.

For the M-score, in the original study, Beneish (1999) used a balance sheet approach to determine total accruals (refer to *TATA_BS* in Appendix 1). However, more recent studies (such as Beneish et al. (2013) and Beneish and Vorst (2021)) have used an income statement approach to determine total accruals (refer to *TATA_IS* in Appendix 1). This change was driven by new disclosure requirements in financial reporting standards (Beneish et al., 2013). This study presents both methods separately, referred to as the M-score (BS) and M-score (IS). In addition, all three cut-off points (i.e. −1.49, −1.78 and −1.89) were used to predict whether an observation was manipulated.

For the F-score, Dechow et al. (2011) determined the UEM to be 0.0037 based on their sample of US companies. However, it is unclear in the literature whether the UEM should be updated for country-specific risk, particularly given SA's higher risk of economic crime (PricewaterhouseCoopers, 2020). As a result, this study used the original US UEM of 0.0037 and a recalculated UEM specific to the SA sample of 0.0098 (23/2343).

3.4. Testing the detective power of the M- and F-scores

Following the estimation of the M- and F-scores, various classification performance metrics and the area under the receiver operating characteristic curve (AUC) were used to test the detective power of models in SA. Mongwe and Malan (2020) identify the common classification performance metrics in the literature as follows:

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + False\ positive + True\ negative + False\ negative} \tag{7}$$

$$Sensitivity = \frac{True\ positive}{True\ positive + False\ negative} \tag{8}$$

$$Specificity = \frac{True\ negative}{True\ negative + False\ positive} \tag{9}$$

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \tag{10}$$

$$F - measure = \frac{2\ x\ Precision\ x\ Sensitivity}{Precision + Sensitivity} \tag{11}$$

While classification accuracy was the most commonly used measure in the prior literature, it is not appropriate due to the scarcity of financial statement fraud cases (Mongwe & Malan, 2020). Instead, sensitivity and precision are superior in such situations (Moepya, 2017). For this study, the accuracy, sensitivity, precision and F-measure are presented for a clearer picture of classification performance.

The final measure of model performance is the AUC. This measure provides a single statistic based on the sensitivity and specificity of the model. A higher AUC statistic represents better model performance, with an AUC of one representing perfect prediction and an AUC of 0.5 representing a random guess.

3.5. Investigating the earnings management characteristics of false positives

This research focused on accruals-based earnings management (AEM) and companies that meet or just beat prior-year earnings to investigate the earnings management characteristics of false-positive observations.

AEM was measured using the cross-sectional modified Jones model. This model is considered one of the most powerful accruals-based models and is widely used throughout the earnings management literature (Mishra & Malhotra, 2016; Rabin, 2016). This model was estimated as follows:

$$NDA_{i,t} = \alpha_1\left(\frac{1}{A_{i,t-1}}\right) + \alpha_2\left(\frac{\Delta REV_{i,t} - \Delta REC_{i,t}}{A_{i,t-1}}\right) + \alpha_3\left(\frac{PPE_{i,t}}{A_{i,t-1}}\right) + \varepsilon_{i,t} \tag{12}$$

Where $NDA_{i,t}$ represents the estimated non-discretionary accruals for company *i* in year *t*, $A_{i,t-1}$ represents total assets for company *i* in the year *t-1*, $\Delta REV_{i,t}$ represents the change in revenue for company *i* between years *t* and *t-1*, $\Delta REC_{i,t}$ is the change in net receivables for company *i* between years *t* and *t-1*, and $PPE_{i,t}$ is the gross property, plant and equipment for company *i* in year *t* (Dechow et al., 1995). The residual from Equation (12) represents the discretionary accrual element. A Wilcoxon rank sum test was used to identify any statistically significant difference in the means of the discretionary accruals between the manipulator and non-manipulator samples.

Earnings per share (EPS) was used to identify companies that meet or just beat the prior year's earnings, defined as the change in EPS falling between zero and a small positive number. For robustness, three measures of a small positive number were used; namely, a one, two or three cents change in EPS (Lo et al., 2017). A Pearson Chi-squared test was used to identify any statistically significant difference between the two samples' proportions of meet or just beat prior year EPS.

3.6. Re-estimating the coefficients of the M- and F-Scores for the SA context

The coefficients of the M- and F-scores were re-estimated by applying the same variables and methodologies (i.e. probit and logit estimation, respectively) originally used by Beneish (1999) and Dechow et al. (2011) to the current SA data. To determine the appropriate cut-offs for the M-score, following Beneish (1999), the ECM was minimised at the cost-error ratios of 10:1, 20:1, 30:1 and 40:1. The ECM was calculated as:

$$ECM = P(M)P_I C_I + [1 - P(M)]P_{II}C_{II} \tag{13}$$

Where *P(M)* represents the prior probability of encountering earnings manipulators (calculated as 0.0098), P_I and P_{II} represent the probability of type I and type II errors, respectively, and C_I and C_{II} represent the relative costs of type I and type II errors respectively (Beneish, 1999). For the F-score, the UEM of 0.0037 and 0.0098 were used with the cut-off of one representing above-average risk observations.

Classification performance and the AUC were used to compare the detective powers of the original models compared to the re-estimated models. For the classification performance metrics, the number of manipulator companies was considered too small to keep a holdout sample.

When determining the AUC, k-fold cross-validation with ten folds was used. Determining the out-of-sample prediction error is essential to avoid hindsight bias when developing predictive

models. K-fold cross-validation is considered superior to bootstrapping procedures, which overlap the training and test samples. This overlap underestimates the prediction error (Witten et al., 2011). Following Larcker and Zakolyukina (2012) and Moepya (2017), ten folds were used. The AUC of the ten iterations were then averaged to determine the overall AUC.

4. Results and discussion

4.1. Descriptive statistics

Table 3 below presents descriptive statistics on the breakdown of manipulated and non-manipulated observations across industries. Although basic materials, consumer services and industrials are the three largest sectors in the SA economy, they only account for a combined total of five (21.74%) manipulated observations. Instead, consumer goods, a medium-sized sector, accounts for sixteen (69.57%) of the manipulated observations. This is due to the companies involved in SA's recent major frauds (i.e. Steinhoff and Tongaat-Hulett) being classified in this sector. SA's three smallest sectors (healthcare, oil and gas, and telecommunications) have no manipulated observations.[6]

Table 3 also presents the average size and return on assets for the manipulated and non-manipulated samples. On average, manipulated observations tend to be smaller and show a lower return on assets. This lower average performance may have provided an incentive for the companies to engage in manipulative practices.

4.2. Distribution of variables underlying the M- and F-scores

Table 4 Panel A presents the distribution of the variables underlying the M-score for manipulators and non-manipulators for the current sample compared to those obtained by Beneish (1999). Unlike in the original study, where a significant difference in mean between manipulators and non-manipulators was found for five of the eight variables, in the current sample, a significant difference was only found for one variable (*TATA_BS*). This finding is also contrary to Marinakis (2011) and Tarjo and Herawati (2015), who found that four of the eight variables could be used to detect manipulation.

Similarly, Panel B shows the distribution of variables underlying the F-score and the comparison to the original study by Dechow et al. (2011). Unlike the original study, for which six of the seven variables showed a significant difference between manipulators and non-manipulators, only the AISS variable showed a significant difference in the current sample. This finding is contrary to

Table 3. Industry classifications and descriptive statistics

	Manipulated observations	Non-manipulated observations	Total
Industry classification	23	2 320	2 343
Basic materials	2	580	582
Consumer goods	16	238	254
Consumer services	2	473	475
Healthcare	0	79	79
Industrials	1	728	729
Oil and gas	0	35	35
Technology	2	143	145
Telecommunications	0	44	44
Size (R'000)	31 665 723.61	35 788 356.25	35 747 886.53
Return on assets (%)	3.52	9.86	9.80

Note: (Source: Researchers' own construction)

Table 4. Distribution of underlying M- and F-score variables for manipulators and non-manipulators

Panel A: Beneish (1999) M-score

	Current sample			Beneish (1999)		
Characteristic	**Manipulators mean**	**Non-manipulators mean**	**Wilcoxon Z**	**Manipulators mean**	**Non-manipulators mean**	**Wilcoxon Z p-value[1]**
DSRI	1.1386	1.0710	−1.2680	1.465	1.031	0.000
GMI	0.8986	1.0044	0.2970	1.193	1.014	0.006
AQI	1.0529	1.1464	0.0970	1.254	1.039	0.096
SGI	1.0944	1.1278	0.5050	1.607	1.134	0.000
DEPI	1.0788	1.0571	−1.5100	1.077	1.001	0.307
SGAI	1.0170	1.1006	−0.1050	1.041	1.054	0.271
LVGI	1.0217	1.0469	−0.3730	1.111	1.037	0.394
TATA_BS	−0.0042	−0.0287	−2.1900**	0.031	0.018	0.000
TATA_IS[2]	−0.0061	−0.0006	0.0910			

Panel B: Dechow et al. (2011) F-score

	Current sample			Dechow et al. (2011)		
Characteristic	Manipulators mean	Non-manipulators mean	Wilcoxon Z[3]	Manipulators mean	Non-manipulators mean	Pairwise t-test
RSST	0.0313	0.0369	−0.4190	0.135	0.044	3.85***
REC	0.0185	0.0173	−0.7230	0.071	0.028	6.12***
INV	0.0079	0.0122	−0.0600	0.046	0.023	4.22***
SASS	0.5925	0.5588	−0.5710	0.647	0.611	3.87***
CSALES	0.0558	0.1245	0.2260	0.466	0.257	3.88***
ROA	−0.0048	−0.0085	0.5930	−0.032	−0.025	−0.38
AISS	0.9565	0.8039	−1.8400*	0.938	0.869	4.75***

Note: [1]Beneish (1999) *did not present the Wilcoxon Z, but only the associated p-value.*

[2]*Total assets to total accruals based on the income statement approach was not used in the original Beneish (1999) study, so the means and p-value are unavailable.*

[3]*Given the nature of the underlying data, a Wilcoxon rank sum test was more appropriate than the pairwise t-test used by Dechow et al. (2011). A t-test did not yield different conclusions to the Wilcoxon rank sum test.*

**Statistical significance <0.10, **Statistical significance <0.05, ***Statistical significance <0.01*

Note: (Source: Researchers' own construction, as well as data obtained from Beneish (1999) and Dechow et al. (2011))

cogent economics & finance

Bertomeu et al. (2021), who found that the variables included in the F-score are influential in detecting manipulation. However, it does align with Deniswara et al. (2022), Hung et al. (2017) and Putra and Dinarjito (2021), who found that the variables underlying the F-score had limited, if any, ability to distinguish between manipulating and non-manipulating companies in Indonesia.

The lack of a statistically significant distribution of the underlying variables indicates that these variables appear unable to differentiate between manipulating and non-manipulating firms in the current SA sample.

4.3. Detective power of the M- and F-scores

The classification performance of the M- and F-scores in SA at various cut-offs and UEMs are summarised in Table 5. The accuracy (i.e. correct classification of both manipulators and non-manipulators) of the M-scores across all cut-offs is high, comparable to the original study. This high accuracy is also in accordance with studies by Aghghaleh et al. (2016), Beneish and Vorst (2021) and Tarjo and Herawati (2015), who report accuracies of 73.17%, 82.59% and 78.57%, respectively.[7] For the F-score, the SA-specific UEM of 0.0098 yields the highest accuracy of all the models. However, the original UEM of 0.0037 produces the lowest accuracy of all the models, which is reasonably in line with the original study results as well as subsequent results of Aghghaleh et al. (2016), Beneish and Vorst (2021) and Chakrabarty et al. (2022), who report accuracy levels of 76.22%, 60.71% and 57.60% respectively[7]. However, the high accuracy across all models benefits from the imbalance between manipulators and non-manipulators. As such, it is primarily driven by the correct classification of the non-manipulator sample (true negatives).

For sensitivity, which measures the scores' ability to classify manipulating firms correctly, the M-score performs poorly: the best variation of the score can identify only 26.09% of manipulators. At all cut-off levels, the results of the current study are substantially worse than the original study, as well as studies by Aghghaleh et al. (2016), Beneish et al. (2013) and Tarjo and Herawati (2015), who reported sensitivity of 69.51%, 71.00% and 77.10% respectively. However, these results align with Beneish and Vorst (2021) and Lu and Zhao (2021), who found that the M-score could only correctly predict 23.18% and 29.63% of manipulators, respectively. For the F-score, sensitivity is also low, with the UEM of 0.0037 achieving the highest sensitivity of 52.17%, which is worse than the original study. The performance of the F-score is also worse than subsequent studies by 2016), Beneish and Vorst (2021) and Chakrabarty et al. (2022) (for the in-sample test), who reported sensitivities of 73.17%, 64.71% and 68.50% respectively. In their out-of-sample test, however, Chakrabarty et al. (2022) reported a sensitivity of 54.61%, which is more in line with the current study.

In terms of precision (i.e. the ability to classify only true manipulators as manipulators) and the F-measure (a metric which combines sensitivity and precision), the M-score's performance in the SA sample is poor compared to what was achieved in the original study as well as studies by Aghghaleh et al. (2016) of 75.00% and 72.15%[7] respectively as well as Tarjo and Herawati (2015) of 79.41% and 78.26%[7] respectively. However, the M-score's precision and F-measure are similar to the results achieved by Beneish and Vorst (2021) of 0.76% and 1.48%[7], respectively. Surprisingly, the F-score (UEM = 0.0037) achieves higher precision and F-measure than the original study, despite being worse at correctly classifying manipulators. Further, the precision and F-measure of the F-score (UEM = 0.0037) are in line with other studies by Beneish and Vorst (2021) and Chakrabarty et al. (2022), who report a precision of 0.92% and 1.13%[7] respectively and an F-measure of 1.81% and 2.22%[7] respectively.

Considering the performance across scores, the M-score (BS) outperforms the M-score (IS) across all metrics for equivalent cut-offs (except for sensitivity and the type I error at the highest cut-off of −1.49, which are equal). By comparison, the F-score (UEM = 0.0098) has the highest accuracy across all scores, while the F-score (UEM = 0.0037) has the lowest accuracy. Despite this low accuracy, the F-score (UEM = 0.0037) is the best-performing score in terms of sensitivity. In

Table 5. Classification performance of the M- and F-scores in the SA context

	M-score (BS) (Cut-off = -1.49)	M-score (BS) (Cut-off = -1.78)	M-score (BS) (Cut-off = -1.89)	M-score (IS) (Cut-off = -1.49)	M-score (IS) (Cut-off = −1.78)	M-score (IS) (Cut-off = −1.89)	F-score (UEM = 0.0098)	F-score (UEM = 0.0037)
Predicted manipulators	**224**	**342**	**408**	**261**	**459**	**553**	**62**	**926**
True positive	1	4	6	1	3	5	0	12
False positive	223	338	402	260	456	548	62	914
Predicted non-manipulators	**2 119**	**2 001**	**1 935**	**2 082**	**1 884**	**1 790**	**2 281**	**1 417**
True negative	2 097	1 982	1 918	2 060	1 864	1 772	2 258	1 406
False negative	22	19	17	22	20	18	23	11
Accuracy	**89.54%**	**84.76%**	**82.12%**	**87.96%**	**79.68%**	**75.84%**	**96.37%**	**60.52%**
Per original estimation sample[1]	*91.41%*[2]	*85.84%*[2]	*82.31%*[2]					*63.71%*
Per original holdout sample[1]	*95.63%*[2]	*92.17%*[2]	*90.38%*[2]					*61.73%*
Sensitivity	**4.35%**	**17.39%**	**26.09%**	**4.35%**	**13.04%**	**21.74%**	**0.00%**	**52.17%**
Per original estimation sample[1]	*58.00%*	*74.00%*	*76.00%*					*68.62%*
Per original holdout sample[1]	*37.50%*	*50.00%*	*56.10%*					*73.83%*
Precision	**0.45%**	**1.17%**	**1.47%**	**0.38%**	**0.65%**	**0.90%**	**0.00%**	**1.30%**
Per original estimation sample[1]	*18.24%*[2]	*13.55%*[2]	*11.28%*[2]					*0.70%*[2]
Per original holdout sample[1]	*13.64%*[2]	*9.30%*[2]	*8.36%*[2]					*1.15%*[2]

(*Continued*)

cogent economics & finance

Table 5. (*Continued*)

	M-score (BS) (Cut-off = -1.49)	M-score (BS) (Cut-off = -1.78)	M-score (BS) (Cut-off = -1.89)	M-score (IS) (Cut-off = -1.49)	M-score (IS) (Cut-off = −1.78)	M-score (IS) (Cut-off = −1.89)	F-score (UEM = 0.0098)	F-score (UEM = 0.0037)
F-measure	**0.87%**	**2.19%**	**2.78%**	**0.70%**	**1.24%**	**1.74%**	**N/A[4]**	**2.53%**
Per original estimation sample[1]	*27.75%[2]*	*22.91%[2]*	*19.64%[2]*					*1.38%[2]*
Per original holdout sample[1]	*20.00%[2]*	*15.69%[2]*	*14.55%[2]*					*2.27%[2]*
Type I error	**95.65%**	**82.61%**	**73.91%**	**95.65%**	**86.96%**	**78.26%**	**100.00%**	**47.83%**
Per original estimation sample[1]	*42.00%[2]*	*26.00%[2]*	*24.00%[2]*					*31.38%*
Per original holdout sample[1]	*62.50%[2]*	*50.00%[2]*	*43.90%[2]*					*26.17%*
Type II error	**9.61%**	**14.56%**	**17.32%**	**11.20%**	**19.65%**	**23.61%**	**2.67%**	**39.38%**
Per original estimation sample[1]	*7.60%*	*13.80%*	*17.50%*					*36.31%*
Per original holdout sample[1]	*3.50%*	*7.20%*	*9.10%*					*38.35%*
Area under the receiver operator curve3	0.5936			0.4789			0.6067	

Note: [1]*The performance measures for the original studies relate to Beneish (1999) for the M-score and Dechow et al. (2011) for the F-score. No original performance measures were presented for the M-score based on the income statement approach or for the F-score based on the UEM of 0.0098, as these were not included in the original study.*

[2]*These performance measures were not presented in the original studies by Beneish (1999) and Dechow et al. (2011), but they have been recalculated based on the data presented in these studies.*

[3]*The results for the area under the receiver operator curve were based on the underlying M- or F-score rather than a specific cut-off point or unconditional expectation of misstatement. The original studies did not include a calculation of the AUC, nor was it possible to recalculate based on the data presented in the studies.*

[4]*Due to both precision and sensitivity being equal to zero, it was impossible to compute the F-measure.*

(Source: Researchers' own construction, as well as data obtained from Beneish (1999) and Dechow et al. (2011)).

cogent economics & finance

addition, it is only outperformed in terms of precision and the F-measure by the M-score (BS) at the lowest cut-off point (−1.89).

The AUC reflects that both the M-score (BS) and F-score outperform a random guess, while the M-score (IS) does not. The AUC for the M-score (BS) of 0.5936 is substantially below Price et al. (2011), who report an AUC of 0.7324, but more in line with the AUC of 0.5770 reported by Beneish and Vorst (2021). The AUC for the F-score of 0.6067 is below that achieved in studies by Beneish and Vorst (2021), Chakrabarty et al. (2022), Price et al. (2011) and Walker (2020) of 0.6730, 0.6670, 0.7238 and 0.6600 respectively. While the F-score slightly outperforms the M-score based on this metric, Price et al. (2011) caution against such an interpretation as the AUC does not distinguish well between two "good" models.

Despite the high overall accuracy of the models, their ability to correctly predict manipulators is low, as shown by the poor sensitivity, precision and type I error metrics. Based on this, hypothesis 1, that the M- and F-scores can detect manipulation in SA, is not supported. Further, while the F-score does outperform the M-score on some metrics, it underperforms on other metrics, depending on the cut-off points used. Thus, there is insufficient evidence to support hypothesis 2, that the F-score outperforms the M-score in the SA context.

4.4. Earnings management characteristics of the false positives

Given the inability of the M- and F-scores to identify manipulators in SA, it is helpful to consider the earnings management characteristics of the false positives to understand better what the models are identifying. Table 6 summarises these results. Panel A compares the false positives to the manipulator sample, whereas Panel B compares the false positives to the true negatives.

For the M-score (BS), the false positive samples do not display similar discretionary and absolute discretionary accruals levels compared to the manipulator sample. Rather, all three cut-offs display higher discretionary and absolute discretionary accruals levels. The F-score (UEM = 0.0098) shows similar results. However, for the M-score (IS) and the F-score (UEM = 0.0037), there is no statistically significant difference between the discretionary accruals and absolute discretionary accruals for the false positive and manipulator samples. For all scores, the discretionary and absolute discretionary accruals of the false positive samples are significantly different from the true negative samples. This indicates that the false positive samples have similar or higher levels of AEM compared to the manipulator sample. It also shows that the false positive samples do not share the same level of AEM compared to the true negative sample.

For all scores, the false positive samples do not display a significantly different proportion of observations that meet or just beat prior year EPS at any level (1, 2 or 3 cents) compared to the manipulators. However, for the M-score models, the false positive samples reveal a higher proportion of observations just beating the prior year's EPS by 1 cent compared to the true negative samples. At the 2 and 3-cent levels, there is no difference between the manipulators, true negatives or false positives for the M-score. For the F-score (UEM = 0.0037), there is a lower proportion of false positives, which just beat the prior year's EPS by 2 and 3 cents compared to the true negatives.

Thus, the evidence presented indicates that the false positives, as determined by the M-score (IS) and F-score (UEM = 0.0037), share similar AEM characteristics as the manipulators, while the false positives, as determined by the M-score (BS) and F-score (UEM = 0.0098), show higher levels of AEM compared to manipulators. When considering earnings thresholds, the M-score (both BS and IS) false positives display similar proportions of meeting or just beating prior year EPS by 1 cent to the manipulators. Considering the F-score, false positives do not display different meet, or just beat, prior year EPS by 1 cent to either the manipulators or the true negatives.

Table 6. Comparison of earnings management characteristics of false positives to true manipulators and non-manipulators

	M-score (BS) (Cut-off = -1.49)	M-score (BS) (Cut-off = -1.78)	M-score (BS) (Cut-off = -1.89)	M-score (IS) (Cut-off = -1.49)	M-score (IS) (Cut-off = -1.78)	M-score (IS) (Cut-off = -1.89)	F-score (UEM = 0.0098)	F-score (UEM = 0.0037)
Panel A: False positives compared to manipulators								
DISCRETIONARY ACCRUALS								
False positive mean	0.0836	0.0732	0.0706	0.0500	0.0454	0.0470	0.1460	0.0145
Manipulator mean	0.0280	0.0280	0.0280	0.0280	0.0280	0.0280	0.0280	0.0280
Wilcoxon Z	2.0060**	2.1000**	2.1690**	1.3940	1.2940	1.2250	3.4410***	−0.0920
ABSOLUTE DISCRETIONARY ACCRUALS								
False positive mean	0.1396	0.1239	0.1157	0.1067	0.0974	0.0959	0.1623	0.0755
Manipulator mean	0.0684	0.0684	0.0684	0.0684	0.0684	0.0684	0.0684	0.0684
Wilcoxon Z	2.3240**	2.241**	2.084**	1.6390	1.4790	1.2600	2.8700***	0.1200
MEET OR JUST BEAT − 1 CENT								
False positive mean	0.0448	0.0414	0.0423	0.0423	0.0417	0.0365	0.0000	0.0186
Manipulator mean	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Pearson Chi^2	1.0751[1]	0.9911[1]	1.0132[1]	1.0124[1]	0.9979[1]	0.8699[1]	N/A	0.4357[1]
MEET OR JUST BEAT − 2 CENT								
False positive mean	0.0583	0.0621	0.0597	0.0538	0.0570	0.0511	0.0192	0.0339
Manipulator mean	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

(*Continued*)

cogent economics & finance

Table 6. (*Continued*)

	M-score (BS) (Cut-off = -1.49)	M-score (BS) (Cut-off = -1.78)	M-score (BS) (Cut-off = -1.89)	M-score (IS) (Cut-off = -1.49)	M-score (IS) (Cut-off = -1.78)	M-score (IS) (Cut-off = -1.89)	F-score (UEM = 0.0098)	F-score (UEM = 0.0037)
Pearson Chi^2	1.4156[1]	1.5173[1]	1.4553[1]	1.3029[1]	1.3867[1]	1.2358[1]	0.4483[1]	0.8068[1]
MEET OR JUST BEAT – 3 CENT								
False positive mean	0.0718	0.0799	0.0746	0.0692	0.0768	0.0693	0.0769	0.0481
Manipulator mean	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Pearson Chi^2	1.7650[1]	1.9858[1]	1.8468[1]	1.7005[1]	1.9045[1]	1.7086[1]	1.8689[1]	1.1618[1]
Panel B: False positives compared to true negatives								
DISCRETIONARY ACCRUALS								
False positive mean	0.0836	0.0732	0.0706	0.0500	0.0454	0.0470	0.1460	0.0145
True negative mean	−0.0105	−0.0142	−0.0164	−0.0079	−0.0128	−0.0162	−0.0044	−0.0115
Wilcoxon Z	9.4700***	12.2510***	13.8000***	8.4250***	11.0160***	12.0470***	6.9380***	7.3400***
ABSOLUTE DISCRETIONARY ACCRUALS								
False positive mean	0.1396	0.1239	0.1157	0.1067	0.0974	0.0959	0.1623	0.0755
True negative mean	0.0637	0.0620	0.0618	0.0666	0.0646	0.0635	0.0692	0.0683
Wilcoxon Z	8.1390***	9.7460***	9.8110***	6.1660***	7.4610***	7.1930***	5.2560***	1.8630*
MEET OR JUST BEAT – 1 CENT								

(*Continued*)

cogent economics & finance

Table 6. (*Continued*)

	M-score (BS) (Cut-off = -1.49)	M-score (BS) (Cut-off = -1.78)	M-score (BS) (Cut-off = -1.89)	M-score (IS) (Cut-off = -1.49)	M-score (IS) (Cut-off = -1.78)	M-score (IS) (Cut-off = -1.89)	F-score (UEM = 0.0098)	F-score (UEM = 0.0037)
False positive mean	0.0448	0.0414	0.0423	0.0423	0.0417	0.0365	0.0000	0.0186
True negative mean	0.0224	0.0217	0.0209	0.0223	0.0204	0.0209	0.0251	0.0284
Pearson Chi^2	4.2315**	4.6879**	6.3707**	3.8446**	6.9231***	4.2590**	1.3398[1]	2.2425
MEET OR JUST BEAT – 2 CENT								
False positive mean	0.0583	0.0621	0.0597	0.0538	0.0570	0.0511	0.0192	0.0339
True negative mean	0.0448	0.0434	0.0433	0.0451	0.0435	0.0446	0.0467	0.0541
Pearson Chi^2	0.8313	2.3050	2.0386	0.3972	1.5319	0.4035	0.8743[1]	5.1057**
MEET OR JUST BEAT – 3 CENT								
False positive mean	0.0718	0.0799	0.0746	0.0692	0.0768	0.0693	0.0769	0.0481
True negative mean	0.0582	0.0560	0.0563	0.0583	0.0553	0.0564	0.0591	0.0669
Pearson Chi^2	0.6635	2.9428*	1.9934	0.4974	3.0263*	1.2469	0.2892[1]	3.4684*

Note: [1]*The statistical significance of the Pearson Chi^2 for these characteristics was based on Fisher's exact rather than the standard p-value due to the small sample size. No differences in the significance level would have arisen had the standard p-value been used.*

****, ** and * represent statistical significance at 1%, 5% and 10% respectively. (Source: Researchers' own construction)*

cogent economics & finance

As a result, hypothesis 3 is partially supported as the false positive samples appear to have similar or higher levels of AEM than the manipulator sample and share similar meet, or just beat, EPS characteristics, but only at the 1 cent level for the M-score.

4.5. Re-estimating the coefficients of the M-Score and F-Score

Due to the poor performance of the original M- and F-scores in detecting manipulation in SA, the models were re-estimated to determine the coefficients that apply in SA, as shown in Table 7. Except for the constant terms and the *AISS* term in the F-score, none of the variable coefficients were statistically significant. This closely mirrors Table 4, where only *TATA_BS* and *AISS* showed a significant difference between manipulators and non-manipulators. These results again revealed the inability of the underlying variables to distinguish between manipulators and non-manipulators in SA.

It should also be noted that the models as a whole display little explanatory power. All revised models have insignificant LR Chi^2 statistics and low pseudo R^2 statistics. This contradicts Marinakis (2011), who used UK data to report a revised M-score model with a statistically significant Chi^2 at the 1% level and a pseudo R^2 of 0.318.

Following the estimation of the models in Table 7, the M-score cut-offs that minimised the ECM were determined at relative costs of 10:1, 20:1, 30:1 and 40:1. Like Beneish (1999), the ECM at the relative costs of 20:1 and 30:1 were the same for both the balance sheet and income statement versions, resulting in the same cut-off. These cut-offs were determined as −1.7910 (10:1), −1.9653 (20:1) and −2.0407 (40:1) for the M-score (BS) and −1.4539 (10:1), −1.9735 (20:1) and −2.1641 (40:1) for the M-score (IS).

Table 8 presents the classification performance for the revised M- and F-score models based on the estimation sample. Comparing the re-estimated models' classification performance to the original models' performance produced mixed results. For the M-score, the revised models performed better than the original models in this sample for accuracy, precision, the F-measure and the type II error. In contrast, the original models performed better in terms of sensitivity and the type I error. Thus, re-estimating the M-score coefficients reduced sensitivity but improved precision. By comparison, for the F-score, the revised scores performed better than the original scores for sensitivity and the type I error. In contrast, the original scores were superior in terms of accuracy and the type II error. The precision and F-measure were comparable and produced mixed results depending on the selected UEM. Thus, re-estimating the F-score improved sensitivity but at the cost of a higher type II error. This trade-off between sensitivity and precision in fraud detection models is also identified by Beneish and Vorst (2021). It should be noted, however, that this comparison for the re-estimated models is based on the estimation sample and may suffer from hindsight bias. Further out-of-sample testing is required to validate these findings, but could not be performed on these data due to the small sample of manipulators.

The AUC results were more robust as they were based on k-fold cross-validation using ten folds. Here, the revised models were consistently outperformed by the original models.

Unfortunately, comparable studies such as Cecchini et al. (2010) and Marinakis (2011), who also re-estimated the coefficients of the M- and F-scores, did not provide comparative results between the original and the re-estimated coefficients. However, studies which added variables to the models before re-estimation have shown improved performance across all metrics. For the M-score, Marinakis (2011) revised model outperformed his re-estimated model for accuracy, sensitivity and precision in both the estimation and holdout samples for all relative cost levels. Likewise, Chakrabarty et al. (2022) revised F-score outperformed the original F-score based on the same metrics as well as the AUC, which increased from 0.6670 to 0.7271.

Table 7. Re-estimated coefficients for the M- and F-scores in the SA context

Panel A: Re-estimated M-score			Panel B: Re-estimated F-score	
	BS	**IS**		
DSRI	0.0776 (0.1622)	0.1194 (0.1506)	**RSST**	−0.4474 (1.7727)
GMI	−0.1463 (0.1497)	−0.1355 (0.1412)	**REC**	0.9622 (3.9236)
AQI	−0.0692 (0.1237)	−0.0722 (0.1267)	**INV**	−2.8220 (5.9092)
SGI	−0.2178 (0.2997)	−0.1280 (0.2819)	**SASS**	0.7275 (0.9417)
DEPI	0.0307 (0.1848)	0.0347 (0.1783)	**CSALES**	−0.5816 (0.7008)
SGAI	−0.2042 (0.2180)	−0.1875 (0.2070)	**ROA**	0.8299 (2.1063)
TATA_BS	1.4912 (1.0488)		**AISS**	1.7646* (1.0284)
TATA_IS		−0.3350 (0.9243)	**Constant**	−6.5314*** (1.1503)
LVGI	−0.0603 (0.2713)	−0.1538 (0.2819)		
Constant	−1.6897*** (0.6040)	−1.7944*** (0.5898)		
Observations	2 343	2 343	**Observations**	2 343
LR Chi^2	5.10	3.19	**LR Chi^2**	6.68
Prob > Chi^2	0.7464	0.9219	**Prob > Chi^2**	0.4631
Pseudo R^2	0.0197	0.0123	**Pseudo R^2**	0.0258

Note: No evidence of multicollinearity was identified. Standard errors are presented in parentheses. The re-estimated M-score coefficients were based on probit estimation, while the re-estimated F-score coefficients were based on logit estimation, in line with the original studies.

**** and * represent statistical significance at 1% and 10%, respectively.*

(Source: Researchers' own construction)

Given the mixed results presented above, hypothesis 4 is partially accepted. The re-estimated M-score failed to improve the identification of manipulators but did reduce misclassification errors. On the other hand, the re-estimated F-score improved the identification of manipulators but failed to reduce misclassification errors.

4.6. Additional tests

Following Beneish (1999), the manipulator sample was matched to the non-manipulators based on industry and year as an additional test. Regarding the classification performance of the original M- and F-scores, the scores' accuracy, precision and F-measure were marginally superior compared to the unmatched results. The matched AUC was also marginally better than the unmatched AUC. The sensitivity, however, remained unchanged. Regarding the earnings management characteristics of the false positives, the results using the matched data revealed no differences compared to using the unmatched data. Finally, regarding the re-estimated M- and F-scores, the comparative performance of the matched data for the M-score was mixed. The matched data results were marginally worse than the unmatched data for accuracy. Precision metrics were generally slightly superior for the matched data, but this depended on the relative cost ratio. Sensitivity was unchanged across the matched and unmatched data. Overall, the conclusions drawn remained unchanged, given the additional testing based on the matched data.

Table 8. Classification performance for re-estimated M- and F-scores (estimation sample)

	M-score (BS) (Cut-off = -1.7910)	M-score (BS) (Cut-off = -1.9653)	M-score (BS) (Cut-off = -2.0407)	M-score (IS) (Cut-off = -1.4539)	M-score (IS) (Cut-off = -1.9735)	M-score (IS) (Cut-off = -2.1641)	F-score (UEM = 0.0098)	F-score (UEM = 0.0037)
Predicted manipulators	**10**	**28**	**63**	**1**	**22**	**105**	**1 347**	**1 889**
True positive	1	2	3	0	1	4	17	22
False positive	9	26	60	1	21	101	1 330	1 867
Predicted non-manipulators	**2 333**	**2 315**	**2 280**	**2 342**	**2 321**	**2 238**	**996**	**454**
True negative	2 311	2 294	2 260	2 319	2 299	2 219	990	453
False negative	22	21	20	23	22	19	6	1
Accuracy	**98.68%**	**97.99%**	**96.59%**	**98.98%**	**98.16%**	**94.88%**	**42.98%**	**20.27%**
Sensitivity	**4.35%**	**8.70%**	**13.04%**	**0.00%**	**4.35%**	**17.39%**	**73.91%**	**95.65%**
Precision	**10.00%**	**7.14%**	**4.76%**	**0.00%**	**4.55%**	**3.81%**	**1.26%**	**1.16%**
F-measure	**6.06%**	**7.84%**	**6.98%**	**N/A[1]**	**4.44%**	**6.25%**	**2.48%**	**2.30%**
Type I error	**95.65%**	**91.30%**	**86.96%**	**100.00%**	**95.65%**	**82.61%**	**26.09%**	**4.35%**
Type II error	**0.39%**	**1.12%**	**2.59%**	**0.04%**	**0.91%**	**4.35%**	**57.33%**	**80.47%**
AUC[2]	**0.5339**			**0.4419**			**0.5546**	

Note: [1]Due to both precision and sensitivity being equal to zero, it was impossible to compute the F-measure.

[2]The AUC was based on 10-fold cross-validation.

(Sources: Researchers' own construction)

cogent economics & finance

Further, Kukreja et al. (2020) argue that the M-score cannot detect every type of misstatement. The same may be true of the F-score. Consequently, the classification sensitivity[8] was recalculated based on the separate categories of manipulators (i.e. FSCA enforcement action, FRIP restatement and relevant qualified audit opinion) based on the original versions of the M- and F-scores. The results are presented in Table 9 below.

Regarding FSCA enforcement actions, the scores performed worse for this category. The M-score (BS and IS) could only correctly classify 7.14% of such actions when the broader cut-offs of −1.78 and −1.89 were selected. The narrower cut-off of −1.49 was unable to classify any enforcement action correctly. Likewise, the F-score (UEM = 0.0098) could also not correctly classify any enforcement action. However, the F-score (UEM = 0.0037) identified 50% of such enforcement actions. The scores appeared to perform better with regard to classification sensitivity for FRIP restatements and qualified audit opinions. The M-score (BS) correctly classified 60% of FRIP restatements when using the more lenient −1.89 cut-off. However, the cut-off of −1.49 failed to classify any FRIP restatement correctly. The M-score (IS) performed worse than the M-score (BS) at the broadest cut-off of −1.89 by only correctly identifying 40% of FRIP restatements. However, it performed better at the more stringent −1.49 cut-off as it identified 20% of FRIP restatements. Again, the F-score (UEM = 0.0098) could not identify any FRIP restatements, while the F-score (UEM = 0.0037) performed the best of all the scores and correctly classified 80% of the FRIP restatements. Finally, regarding the qualified audit opinions, the M-score (BS and IS) performed moderately at the broadest cut-off of −1.89, identifying 50% of qualified opinions. At the most stringent cut-off (−1.49), the M-score (BS) outperformed the M-score (IS) but was still only able to identify 25% of qualified opinions. The F-score performed worst of the scores in correctly classifying qualified opinions, only correctly identifying 25% when using the UEM of 0.0037.

These results show that the F-score (UEM = 0.0037) outperformed both M-score models for FSCA enforcement actions and FRIP restatements. However, the M-score outperformed the F-score when identifying qualified audit opinions. Caution, however, should be applied when relying on this set of additional results. Firstly, the sensitivity was based on very few observations, particularly for FRIP restatements and qualified audit opinions. Secondly, only the classification sensitivity is provided. As the false positives would have changed only slightly, the scores would continue to perform poorly in terms of precision, the F-measure and the type I error.

5. Summary of results

This study tested four hypotheses. The first hypothesis, of whether the M- and F-scores could detect financial statement manipulation in SA, was not supported. The second hypothesis was that, based on the findings of prior studies, the F-score would outperform the M-score. Given the inability of both models to successfully detect manipulation in SA, this hypothesis was also not supported. Partial support was found for the third hypothesis, which expected the false positive sample to share similar earnings management characteristics with the manipulator sample. Here, the study found that the false positives tended to have similar or higher levels of discretionary accruals in comparison to the manipulators. Finally, the fourth hypothesis expected that updating the coefficients of the M- and F-scores would improve the models' ability to identify manipulators in SA. This hypothesis was partially supported as, for the M-score, misclassifications were reduced, although the ability to identify manipulators worsened. The opposite occurred with the F-score as correctly identifying manipulators improved, but misclassifications increased substantially. In summary, the findings failed to support hypotheses 1 and 2, while partial support was found for hypotheses 3 and 4.

5.1. Discussion

The performance classification and AUC results reveal that both the M- and F-score appear ineffective in accurately identifying cases of manipulation in the SA context. This is consistent with more recent studies such as Beneish and Vorst (2021), Comporek (2020) and Lu and Zhao (2021), who also found limited ability of the models to detect fraud.

Table 9. Classification sensitivity per manipulator category

SCORE	CATEGORIES OF MANIPULATORS		
	FSCA enforcement action	FRIP restatement	Qualified audit opinion
M-score (BS)			
Cut-off = −1.49	0.00%	0.00%	25.00%
Cut-off = −1.78	7.14%	20.00%	50.00%
Cut-off = −1.89	7.14%	60.00%	50.00%
M-score (IS)			
Cut-off = −1.49	0.00%	20.00%	0.00%
Cut-off = −1.78	7.14%	20.00%	25.00%
Cut-off = −1.89	7.14%	40.00%	50.00%
F-score			
UEM = 0.0098	0.00%	0.00%	0.00%
UEM = 0.0037	50.00%	80.00%	25.00%

One possible explanation is that the models are inappropriate in the SA context. This could be a result of the underlying variables being unable to distinguish between manipulators and non-manipulators, as seen in section 4.2. This explanation is consistent with Lu and Zhao's (2021) argument that the M-score does not work in the Chinese context, given the period it was developed and the different reporting contexts. Such an argument also applies to SA as the period under consideration is predominantly post the 2008 financial crisis. Also, SA is an emerging market and uses IFRS rather than US GAAP.

Further supporting the above explanation are the earnings management characteristics. The false positive sample has either similar or higher levels of AEM than the manipulator sample. In addition, the false positive sample displays higher levels of AEM than the true negative sample. This presents evidence that the M- and F-score may identify firms with high AEM levels. However, Enomoto et al. (2015) claim that SA companies are less likely to manage earnings through AEM and more likely to manage them through real earnings management. Also, the false positive sample shows different proportions of meeting or just beating the prior year's EPS to the true negative sample. However, Pududu and De Villiers (2016) contend that SA may focus on thresholds other than earnings. Thus, models that distinguish between manipulators and non-manipulators based on AEM and earnings thresholds may be inappropriate in SA.

The final support for the M- and F-scores being inappropriate in SA is that the models cannot identify the type of manipulation that occurs in SA. Kukreja et al. (2020) note that different models have different limitations. In particular, the M-score is unable to detect every form of manipulation. This is evident in the SA context from the additional tests where the models show different abilities to detect FSCA enforcement actions, FRIP restatements and qualified opinions. In particular, the M-score appears to struggle with FSCA enforcement actions, while the F-score has the worst performance for qualified audit opinions.

An alternative explanation could be that SA regulators are unable to identify manipulators. In SA, 59% of companies experiencing fraud do not report the fraud to the board, 66% do not report fraud to an appropriate regulator, and 72% do not report to the external auditor (PricewaterhouseCoopers, 2020). This culture of not reporting fraud, together with regulators lacking appropriate resources, lower legal enforcement associated with emerging economies and SA investors being unable to detect earnings management (Rabin, 2016), makes it difficult for regulators to investigate fraudulent activities and take appropriate action. As a result, the models may identify valid manipulators that regulators have not yet identified.

Both explanations provide reasons why re-estimating the coefficients of the original models would be insufficient to improve the ability of the models to detect manipulation without substantially increasing the extent of false positives.

6. Conclusion

This study investigated the ability of two popular fraud detection models (the Beneish (1999) M-score and the Dechow et al. (2011) F-score) to identify manipulating companies in the SA context correctly. Based on a sample of 23 manipulators and 2 320 non-manipulators from 2006 to 2018, the study found that both models showed limited ability to classify manipulators correctly. Further investigation into the earnings management characteristics of the false positive sample revealed that the models might be categorising companies based on AEM and earnings thresholds. While extensive earnings management is associated with financial statement fraud, it is not a definite indication that such fraud is occurring. Finally, updating the coefficients of the two models did improve aspects of detection, but at the cost of another. For example, re-estimating the M-score coefficients generally improved precision but at the expense of sensitivity. Conversely, re-estimating the F-score improved sensitivity but at the cost of an increased type II error. These results indicate that either the models are not appropriate in the SA context or that SA regulators cannot identify manipulators due to a lack of reporting fraudulent activities, a lack of resources and weak legal enforcement.

This study makes several contributions. First, the study investigates the ability of two popular models in fraud detection to identify manipulators in the SA context accurately. The results indicate that stakeholders should apply caution in using such models to predict fraudulent financial reporting, given their inability to accurately classify manipulators without generating many false positives. Additionally, regulators should allocate more resources to identify and combat fraudulent financial reporting. Second, the study provides a caution to other academics. Researchers need to report on a wide range of performance metrics so that users understand what the model does well compared to what it does not do well. In addition, academics, particularly in African contexts, are cautioned against indiscriminately using these models as proxies for fraud risk without extensively testing them in the local context. Third, the study contributes to the academic literature by investigating the earnings management characteristics of false positives generated by the models, showing that the models tend to differentiate companies with high levels of earnings management rather than companies which commit fraud. Finally, the study contributes to the development of fraud detection models by showing that re-estimating the model coefficients is likely insufficient to improve the models' performance, particularly if the underlying variables appear incapable of distinguishing between manipulators and non-manipulators. Instead, the focus should be placed on incorporating new variables that better distinguish between manipulators and non-manipulators, especially as the global economy changes and new reporting conventions and standards are developed.

This research has some limitations that provide avenues for future research. The study investigated the fraud detection ability of only two popular models (which only incorporate information directly obtainable from the financial statements) in an SA-specific context. Consequently, the results may not be generalisable to other countries, even in Africa. Future researchers should test the models' performance in their country's context and employ more sophisticated models (which include non-financial information such as the modified M-score by Lu and Zhao's (2021) and models 2 and 3 of the F-score) and compare their performance. A second limitation is that this study considered only AEM and earnings thresholds when investigating the earnings management characteristics of the false positives. Given that companies may use different types of earnings management to achieve the same goals, future studies may consider investigating the real earnings management characteristics of the false positives as well as identifying other thresholds that may be more applicable in SA. A third limitation of this study is that it only updated the original model coefficients for SA. The study did not attempt to add additional explanatory variables or remove insignificant variables from the models. Subsequent studies should attempt to identify new variables that are superior in discriminating between manipulators and non-manipulators and include such variables when revising such models.

Funding
The University of KwaZulu-Natal University Capacity Development Grant supported this work.

Author details
Alastair Marais[1]
E-mail: maraisa@ukzn.ac.za
ORCID ID: http://orcid.org/0000-0001-7844-278X
Claire Vermaak[1]
ORCID ID: http://orcid.org/0000-0001-6355-3321
Patricia Shewell[1]
ORCID ID: http://orcid.org/0000-0003-0969-7847
[1] School of Accounting, Economics and Finance; University of KwaZulu-Natal, Durban, KwaZulu-Natal, South Africa.

Citation information
Cite this article as: Predicting financial statement manipulation in South Africa: A comparison of the Beneish and Dechow models, Alastair Marais, Claire Vermaak & Patricia Shewell, *Cogent Economics & Finance* (2023), 11: 2190215.

Notes

1. The cost of type I and type II errors cannot be objectively measured. As such, Beneish (1999) used relative costs between the error types to determine the expected cost of misclassification based on the expected loss in value upon the discovery that a company is manipulating its financial statements compared to the appreciation in value of a non-manipulator.
2. Although there are numerous techniques to measure total accruals, when calculating the F-score, this study used the method set out by Richardson et al. (2005) as this was the technique used in the development of the original F-score (Dechow et al., 2011). The calculation of total accruals using Richardson et al. (2005) method can be found in Appendix 2.
3. Soft assets refer to those assets that are neither cash nor property, plant and equipment (Dechow et al., 2011).
4. The difference in the number of fraudulent and non-fraudulent observations between the M- and F-scores is due to missing data.
5. The terms positive and negative refer to the classifications of manipulated and non-manipulated observations respectively. A true positive is when a manipulated observation is correctly classified by the model. A true negative is when a non-manipulated observation is correctly classified by the model. A false positive occurs when a non-manipulated observation is incorrectly classified as a manipulated observation (i.e. a type II error) and, finally, a false negative is when a manipulated observation is incorrectly classified as a non-manipulated observation (i.e. a type I error).
6. In order to account for some sectors and years not having any manipulated observations, additional tests using matched observations (based on industry and year) are performed in section 4.6.
7. These studies did not report this classification performance metric. However, they provided sufficient information for the metric to be recalculated.
8. For this additional test, the study focused on sensitivity as the misclassification of non-manipulators would remain largely unchanged meaning that the measures of accuracy, precision, F-measure and type I and II errors would remain largely unchanged.

Disclosure statement
No potential conflict of interest was reported by the authors.

References

Ackerman, S., Raz, O., Zalmanovici, M., & Zlotnick, A. 2019. Automatically detecting data drift in machine learning classifiers. *Engineering Dependable and Secure Machine Learning Systems workshop at AAAI 2019 conference*. Honolulu, Hawaii.

Aghghaleh, S. F., Mohamed, Z. M., & Rahmat, M. M. (2016). Detecting financial statement frauds in Malaysia: Comparing the abilities of Beneish and Dechow models. *Asian Journal of Accounting and Governance*, *7*(2016), 57–65. https://doi.org/10.17576/AJAG-2016-07-05

Albizri, A., Appelbaum, D., & Nicholas, R. (2019). Evaluation of financial statements fraud detection research: A multi-disciplinary analysis. *International Journal of Disclosure and Governance*, *16*(4), 206–241. https://doi.org/10.1057/s41310-019-00067-9

Association of Certified Fraud Examiners. 2020. *Report to the nations: 2020 global study on occupational fraud and abuse* . Available [Accessed 22 November. 2021]: https://acfepublic.s3-us-west-2.amazonaws.com/2020-Report-to-the-Nations.pdf

Association of Certified Fraud Examiners. 2021. *The next normal: Preparing for a post-pandemic fraud landscape* [Online]. Available [Accessed 14 January. 2022]: https://www.acfe.com/uploadedFiles/ACFE_Website/Content/covid19/Covid-19-Preparing-for-a-Post-Pandemic-Fraud-Landscape.pdf

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research*, *58*(1), 199–235. https://doi.org/10.1111/1475-679X.12292

Beneish, M. D. (1999). The detection of earnings manipulation. *Financial Analysts Journal*, *55*(5), 24–36. https://doi.org/10.2469/faj.v55.n5.2296

Beneish, M. D., Lee, C. M. C., & Nichols, D. C. (2013). Earnings manipulation and expected returns. *Financial Analysts Journal*, *69*(2), 57–82. https://doi.org/10.2469/faj.v69.n2.1

Beneish, M. D., & Vorst, P. 2021. The cost of fraud prediction errors. *The Accounting Review*. Advance online publication: [Accessed 19 July 2022] https://doi.org/10.2308/TAR-2020-0068

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. *Review of Accounting Studies*, *26*(2), 468–519. https://doi.org/10.1007/s11142-020-09563-8

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting management fraud in public companies. *Management Science*, *56*(7), 1146–1160. https://doi.org/10.1287/mnsc.1100.1174

Chakrabarty, B., Moulton, P. C., Pugachev, L., & Wang, X. 2022. Catch me if you can: In search of accuracy, scope, and ease of fraud prediction. *SSRN working paper*. Available[Accessed 19 July 2022]: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3352667

Comporek, M. (2020). The effectiveness of the Beneish model in the detection of accounting violations – the example of companies sanctioned by the Polish Financial Supervisory Authority. *Prace Naukowe Uniwersytetu Ekonomicznego We Wroclawiu*, *64*(10), 18–30. https://doi.org/10.15611/pn.2020.10.02

Dechow, P. M., Ge, W., Larson, C., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research*, *28*(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting earnings management. *The Accounting Review*, *70*(2), 193–225.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1996). Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the SEC. *Contemporary Accounting Research*, *13* (1), 1–36. https://doi.org/10.1111/j.1911-3846.1996.tb00489.x

Deniswara, K., Kesuma, J. T., & Louis, V. 2022. Forensic accounting on Indonesian energy sector with Beneish M-score model analysis. *ICEEEG'22: Proceedings of the 6thInternational Conference on E-Commerce, E-Business and E-Government*. Plymouth, United Kingdom.

Enomoto, M., Kimura, F., & Yamaguchi, T. (2015). Accrual-based and real earnings management: An international comparison for investor protection. *Journal of Contemporary Accounting & Economics*, *11*(3), 183–198. https://doi.org/10.1016/j.jcae.2015.07.001

Herawati, N. (2015). Application of Beneish M-Score models and data mining to detect financial fraud. *Procedia - Social and Behavioral Sciences*, *211*(2015), 924–930. https://doi.org/10.1016/j.sbspro.2015.11.122

Hlobo, M., Moloi, T., & Marx, B. (2022). Framework for screening and evaluating the competencies and qualities of the board of directors in South Africa's stated-owned companies. *Journal of Risk and Financial Management*, *15*(11), 1–15. https://doi.org/10.3390/jrfm15110492

Hung, D. N., Ha, H. T. V., & Binh, D. T. (2017). Application of F-Score in predicting fraud, errors: Experimental research in Vietnam. *International Journal of Accounting and Financial Reporting*, *7*(2), 303–322. https://doi.org/10.5296/ijafr.v7i2.12174

Jones, K. L., Krishnan, G. V., & Melendrez, K. D. (2008). Do models of discretionary accruals detect actual cases of fraudulent and restated earnings? An empirical analysis. *Contemporary Accounting Research*, *25*(2), 499–531. https://doi.org/10.1506/car.25.2.8

Kamal, M. E. M., Salleh, M. F. M., & Ahmad, A. (2016). Detecting financial statement fraud by Malaysian public listed companies: The reliability of the Beneish M-score model. *Jurnal Pengurusan*, *46*(2016), 23–32. https://doi.org/10.17576/pengurusan-2016-46-03

Karpoff, J. M., Koester, A., Lee, D. S., & Martin, G. S. (2017). Proxies and databases in financial misconduct research. *The Accounting Review*, *92*(6), 129–163. https://doi.org/10.2308/accr-51766

Koornhof, C., & Du Plessis, D. (2000). Red flagging as an indicator of financial statement fraud: The perspective of investors and lenders. *Meditari Accountancy Research*, *8*(1), 69–93. https://doi.org/10.1108/10222529200000005

Kukreja, G., Gupta, S. M., Sarea, A. M., & Kumaraswamy, S. (2020). Beneish M-score and Altman Z-score as a catalyst for corporate fraud detection. *Journal of Investment Compliance*, *21*(4), 231–241. https://doi.org/10.1108/JOIC-09-2020-0022

Larcker, D. F., & Zakolyukina, A. A. (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, *50*(2), 495–540. https://doi.org/10.1111/j.1475-679X.2012.00450.x

Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. *Journal of Accounting and Economics*, *63*(1), 1–25. https://doi.org/10.1016/j.jacceco.2016.09.002

Lu, W., & Zhao, X. (2021). Research and improvement of fraud identification model of Chinese A-share listed companies based on M-score. *Journal of Financial Crime*, *28*(2), 566–579. https://doi.org/10.1108/JFC-12-2019-0164

Marinakis, P. 2011. *An investigation of earnings management and earnings manipulation in the UK*. Unpublished PhD thesis, University of Nottingham.

Mavengere, K. (2015). Predicting corporate bankruptcy and earnings manipulation using the Altman Z-score and Beneish M-score. The case of manufacturing firm in Zimbabwe. *International Journal of Management Sciences and Business Research*, *4*(10), 8–14.

Mishra, M., & Malhotra, A. K. (2016). Earnings management practices in Indian companies: A cross-sectional analysis. *Journal of Modern Accounting and Auditing*, *12*(6), 295–305. https://doi.org/10.17265/1548-6583/2016.06.001

Moepya, S. O. 2017. *Enhancing the detection of financial statement fraud through the use of missing value estimation, multivariate filter feature selection and cost-sensitive classification*. Unpublished PhD thesis, University of Johannesburg.

Moepya, S. O., Akhoury, S. S., & Nelwamondo, F. V. 2014. Applying cost-sensitive classification for financial statement fraud detection under high class-imbalance. *2014 IEEE International Conference on Data Mining Workshop*. Shenzhen, China. 183–192.

Moepya, S. O., Akhoury, S. S., Nelwamondo, F. V., & Twala, B. (2016). The role of imputation in detecting fraudulent financial reporting. *International Journal of Innovative Computing, Information and Control*, *12*(1), 333–356.

Moepya, S. O., Nelwamondo, F. V., & Van Der Walt, C. (2014). A support vector machine approach to detect financial statement fraud in South Africa: A first look. In N. T. Nguyen, B. Attachoo, B. Trawinski, & K. Somboonviwat (Eds.), *Intelligent Information and Database Systems. ACIIDS 2014. Lecture Notes in Computer Science* (pp. 42–51). Springer.

Mongwe, W. T., & Malan, K. M. (2020). A survey of automated financial statement fraud detection with relevance to the South African context. *South African Computer Journal*, *32*(1), 74–112. https://doi.org/10.18489/sacj.v32i1.777

Muzata, T., & Marozva, G. (2022). Executive compensation schemes: Accelerants of agency and corporate governance problems in South Africa. *African Journal of Governance and Development*, *11*(1.2), 328–350. https://doi.org/10.36369/2616-9045/2022/v11si2a7

Nyakarimi, S. (2022). Probable earning manipulation and fraud in banking sector. Empirical study from East Africa. *Cogent Economics & Finance*, *10*(1), 1–20. https://doi.org/10.1080/23322039.2022.2083477

Nyakarimi, S. N., Kariuki, S. N., & Kariuki, P. W. (2020). Financial statement manipulations using Beneish model and probit regression model: A case of banking sector in Kenya. *European Online Journal of Natural and Social Sciences*, *9*(1), 253–264.

Okiro, K., & Otiso, D. O. (2021). Detection of fraud in financial statements using Beneish ratios for companies listed at Nairobi Securities Exchange. *African Development Finance Journal*, *5*(1), 92–126.

Onyebuchi, O. H., & Nkem, A. T. (2021). Forensic accounting and economic crime: A study of selected conglomerate firms in Nigeria. *Advance Journal of Management, Accounting and Finance*, *6*(7), 37–67.

Orazalin, N., & Akhmetzhanov, R. (2018). Earnings management, audit quality, and cost of debt: Evidence from a Central Asian economy. *Managerial Auditing Journal*, *34*(6), 696–721. https://doi.org/10.1108/MAJ-12-2017-1730

Paolone, F., & Magazzino, C. (2015). Earnings manipulation among the main industrial sectors. Evidence from Italy. *Economia Aziendale Online*, *5*(4), 253–261.

Perols, J. L., & Lougee, B. A. (2011). The relationship between earnings management and financial statement fraud. *Advances in Accounting, Incorporating Advances in International Accounting*, *27*(1), 39–53. https://doi.org/10.1016/j.adiac.2010.10.004

Price, R. A., Sharp, N. Y., & Wood, D. A. (2011). Detecting and predicting accounting irregularities: A comparison of commercial and academic risk measures. *Accounting Horizons*, *25*(4), 755–780. https://doi.org/10.2308/acch-50064

PricewaterhouseCoopers. 2020. *Economic crime - When the boardroom becomes the battlefield*. Available [Accessed 17 January. 2022]: https://www.pwc.co.za/en/assets/pdf/global-economic-crime-survey-2020.pdf

Pududu, M. L., & De Villiers, C. (2016). Earnings management through loss avoidance: Does South Africa have a good story to tell? *South African Journal of Economic and Management Sciences*, *19*(1), 18–34. https://doi.org/10.4102/sajems.v19i1.1124

Putra, A. N., & Dinarjito, A. (2021). The effect of fraud pentagon and F-score model in detecting fraudulent financial reporting in Indonesia. *Jurnal Ilmiah Akuntansi dan Bisnis*, *16*(2), 247–263. https://doi.org/10.24843/JIAB.2021.v16.i02.p05

Rabin, C. E. 2016. *Earnings management in South Africa: Evidence and implications*. Unpublished PhD thesis, University of the Witwatersrand.

Rad, M., Amiri, A., Ranjbar, M. H., & Salari, H. (2021). Predictability of financial statements fraud-risk using Benford's Law. *Cogent Economics & Finance*, *9*(1), 1–27. https://doi.org/10.1080/23322039.2021.1889756

Richardson, S. A., Sloan, R. G., Soliman, M. T., & Tuna, İ. (2005). Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics*, *39* (3), 437–485. https://doi.org/10.1016/j.jacceco.2005.04.005

Rossouw, J., & Styan, J. (2019). Steinhoff collapse: A failure of corporate governance. *International Review of Applied Economics*, *33*(1), 163–170. https://doi.org/10.1080/02692171.2019.1524043

Skousen, C. J., & Twedt, B. J. (2009). Fraud score analysis in emerging markets. *Cross Cultural Management: An International Journal*, *16*(3), 301–316. https://doi.org/10.1108/13527600910977373

Van Der Linde, K. E. (2022). The Steinhoff corporate scandal and the protection of investors who purchased shares on the secondary market. *Potchefstroom Electronic Law Journal*, *25*(1), 1–23. https://doi.org/10.17159/1727-3781/2022/v25i0a14876

Walker, S. 2020. A needle found: Machine learning does not significantly improve corporate fraud detection beyond a simple screen on sales growth. *SSRN working paper*. Available[Accessed 20 November 2021]: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3739480

Watson, S., & Rossouw, J. (2012). JSE efficiency and share price reaction to forced financial restatements. *Journal of Economic and Financial Sciences*, *5*(2), 417–436. https://doi.org/10.4102/jef.v5i2.292

Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2016). Characterizing concept drift. *Data Mining and Knowledge Discovery*, *30*(4), 964–994. https://doi.org/10.1007/s10618-015-0448-4

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

World Bank. 2020. *Gross domestic product 2020*. Available: https://databank.worldbank.org/data/download/GDP.pdf [Accessed 17 January. 2022].

World Economic Forum. 2017. The global competitiveness report 2016-2017. Available[Accessed 3 August. 2017]: http://www3.weforum.org/docs/GCR2016-2017/05FullReport/TheGlobalCompetitivenessReport2016-2017_FINAL.pdf

Appendix 1.

Beneish (1999) M-Score

The Beneish (1999) M-score is calculated as follows:

$$M = -0.480 + 0.920DSRI + 0.528GMI + 0.404AQI + 0.892SGI + 0.115DEPI - 0.172SGAI + 4.679TATA - 0.327LVGI$$

Where the independent variables are defined and calculated as follows:

Variable	Description	Calculation
DSRI	Day's sales receivable index	$\frac{Accounts\ Receivable_t/Revenue_t}{Accounts\ Receivable_{t-1}/Revenue_{t-1}}$
GMI	Gross margin index	$\frac{(Revenue_{t-1}-Cost\ of\ sales_{t-1})/Revenue_{t-1}}{(Revenue_t-Cost\ of\ sales_t)/Revenue_t}$
AQI	Asset quality index	$\frac{1-((Current\ assets_t+PPE_t)/Total\ assets_t)}{1-((Current\ assets_{t-1}+PPE_{t-1})/Total\ assets_{t-1})}$
SGI	Sales growth index	$\frac{Revenue_t}{Revenue_{t-1}}$
DEPI	Depreciation index	$\frac{Depreciation_{t-1}/(Depreciation_{t-1}+PPE_{t-1})}{Depreciation_t/(Depreciation_t+PPE_t)}$
SGAI	Sales, general and administrative expenses index	$\frac{Sales,\ general\ and\ admin\ expenses_t/Revenue_t}{Sales,\ general\ and\ admin\ expenses_{t-1}/Revenue_{t-1}}$
TATA_BS	Total accruals (based on the balance sheet approach) to total assets	$\frac{\left(\begin{array}{c}\Delta Current\ assets_t - \Delta Cash_t - \Delta Current\ liabilities_t - \\ \Delta Current\ maturities\ of\ long\ term\ debt_t - \\ \Delta Income\ tax\ payable_t - Depreciation\ and\ amortisation_t\end{array}\right)}{Total\ assets_t}$
TATA_IS	Total accruals (based on the income statement approach) to total assets	$\frac{(Income\ before\ extraordinary\ items_t - Cash\ from\ operations_t)}{Total\ assets_t}$
LVGI	Leverage index	$\frac{(Long\ term\ debt_t+Current\ liabilities_t)/Total\ assets_t}{(Long\ term\ debt_{t-1}+Current\ liabilities_{t-1})/Total\ assets_{t-1}}$

Appendix 2: Dechow et al. (2011) F-Score

The Dechow et al. (2011) F-score is calculated as follows:

$$Predicted\ value = -7.893 + 0.790RSST + 2.518\Delta REC + 1.191\Delta INV + 1.979SASS + 0.171\Delta CSALES - 0.932\Delta ROA + 1.029AISS$$

Where the independent variables are defined and calculated as follows:

Variable	Description	Calculation
RSST	Accruals as measured by Richardson et al. (2005)	$\frac{\Delta WC+\Delta NCO+\Delta FIN}{Averagetotalassets}$ Where: $\Delta WC = (Current\ assets\ - Cash\ and\ short\ term\ investments) - (Current\ liabilities\ - Debt\ in\ current\ liabilities)$ $\Delta NCO = (Total\ assets - Current\ assets - Investments\ and\ advances) - (Total\ liabilities - Current\ liabilities - Longtermdebt)$ $\Delta FIN = (Short\ term\ investments + Long\ term\ investments) - (Long\ term\ debt + Debt\ in\ current\ liabilities + Preferred\ stock)$
ΔREC	Change in receivables	$\frac{\Delta\ Accounts\ receivable}{Average\ total\ assets}$
ΔINV	Change in inventory	$\frac{\Delta\ Inventory}{Average\ total\ assets}$
SASS	Percentage soft assets	$\frac{Total\ assets_t - PPE_t - Cash\ and\ cash\ equivalents_t}{Total\ assets_t}$
ΔCSALES	Change in cash sales	Percentage change in cash sales, where cash sales is measured as $(Sales - \Delta Accounts\ receivable)$
ΔROA	Change in return on assets	$\frac{Earnings_t}{Average\ total\ assets_t} - \frac{Earnings_{t-1}}{Average\ total\ assets_{t-1}}$
AISS	Actual issuance	Indicator variable coded to 1 if the firm issued securities during year t.

***Cogent Economics & Finance* (ISSN: 2332-2039) is published by Cogent OA, part of Taylor & Francis Group.**