EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

Econ Journal Watch 19(2)
September 2022: 190–203

Erroneous Erratum to Accounting Fraud Article

Stephen Walker

Link to Abstract

There was a paper in the *Journal of Accounting Research* (Bao et al. 2020) that I criticized in *Econ Journal Watch* (Walker 2021a). The authors published a concurrent reply (Bao et al. 2021), to which I rejoined in the next issue (Walker 2021b) without a second reply by the authors, though they were invited to supply one.

In August 2022, the *Journal of Accounting Research* posted an erratum (Bao et al. 2022) to Bao et al. (2020). Two lines in that erratum jumped out at me. First:

> Walker (2021a and 2021b) identified an error **in the program codes** of Bao et al. (2020) posted on Github that led to an overstatement of model performance metrics. (Bao et al. 2022, 1635, boldface added)

The part in boldface is false. I discovered no errors in their program codes. Rather, the "error" was in their dataset—specifically, how they identified fraud cases in their sample. A coding error may be accidental, but the method used to define their fraud cases was deliberate. To this date, the authors have offered no explanation as to why they did what they did, nor offer any explanation as to what prompted their misidentifying of fraud cases. Their readers deserve an explanation as to the nature and causes of their misidentification.

The second portion of the erratum that jumped out at me is:

> It continues to dominate the performance of the other models for the test period **2003–2005**. Bao et al. (2020) argued that this test period was the cleanest because many accounting frauds in the test years after **2005** could be undetected due to reduced regulatory enforcement of accounting fraud that approximately coincided with the **2008 financial crisis** (see also Donelson et al. 2021). (Bao et al. 2022, 1636, boldface added)

This is also false, and that falseness is easily verifiable from their original 2020 publication:

> **We end the sample in 2008** because there is a noticeable shift in the regulators' enforcement of accounting fraud that approximately coincided with the **2008 financial crisis**. (Bao et al. 2020, 208)

In this paper, I make the case that there is evidence of academic misconduct and make the recommendation that the *Journal of Accounting Research* launch a full and independent investigation into the matter.

Replication of erratum code

First, for completeness, I want to replicate their results posted in their erratum. I downloaded the new code from Github and ran it as is. I also ran a second scenario that optimized the number of tree parameters based on a grid search, which was given to readers in this new version. Recall that the original paper gave no guidance as to how they performed this search. In their reply to me, Bao et al. (2021, 72–73) led with the argument that I had not optimized this tree parameter using grid search. So, I ran this code first and the results are shown in Table 1.

TABLE 1. Grid search results

Trees	AUC	Trees	AUC
100	0.7350	800	0.7490
200	0.7432	900	0.7488
300	0.7451	1000	0.7495
400	0.7459	1500	0.7478
500	0.7446	2000	0.7464
600	0.7470	2500	0.7430
700	0.7484	3000	0.7418

What is obvious is how little this grid search matters in terms of AUC performance. That is a point I had made in my rejoinder, prior to the new Erratum (Walker 2021b, 232). The result of this search shows that optimal number of trees is 1,000. Strangely, the number of trees used in the updated code was 300. Therefore, for this replication, I run two scenarios, one with the code as-is and one with the tree parameter changed to 1,000. I report metrics AUC, NDCG@k (which the authors argued was the most important metric), and the number of correct hits identified by the model, which is the easiest metric to describe. In addition to averaging (which the authors show only), I report results by year for their original primary test period 2003–2008. The results of this analysis are shown in Table 2.

TABLE 2. Summary of performance metrics by year

RUSBOOST—28 Variables									
	AUC			NDCG@1%			Hits (Top 1%)		
Year	Erratum	Run (1)	Run (2)	Erratum	Run (1)	Run (2)	Erratum	Run (1)	Run (2)
2003		0.7281	0.7331		0.0624	0.0654		5	5
2004		0.7646	0.7670		0.0279	0.0269		2	2
2005		0.7289	0.7245		0.0141	0.0148		1	1
2006		0.7590	0.7602		0.0000	0.0000		0	0
2007		0.7176	0.7184		0.0000	0.0000		0	0
2008		0.6280	0.6346		0.0000	0.0000		0	0
Summary									
2003–2005	0.7428	0.7405	0.7415	0.0394	0.0348	0.0357	9	8	8
2006–2008		0.7015	0.7044		0.0000	0.0000	1	0	0
2003–2008	0.7228	0.7210	0.7230	0.0237	0.0174	0.0179	10	8	8

As to the hit number, the authors reported 9 hits from 2003–2005 and 10 hits total from 2003–2008, implying only 1 additional hit from 2006–2008.

I was unable to generate the exact results from the erratum. Using the optimal tree setting, my run shows 8 hits from 2003–2005 and zero hits thereafter, so I have 20 percent fewer hits *using their code*. Since this value represents the same number of hits from the logit model from previous literature shown in Table 1 Panel B of their erratum (Bao et al. 2022, 1637), it would appear that this machine learning approach does not outperform existing models. Recall that the authors originally wrote that:

> **We find that our best model, the ensemble learning model, identified a total of 16 fraud cases in the test period 2003–2008. In contrast, the comparable figure is 9 for the Dechow et al. model and 7 for the Cecchini et al. model.** These results suggest that the differences in the performance of the ensemble learning model versus the two benchmark models are also economically significant. (Bao et al. 2020, 223, boldface added)

So, their paper originally claimed a near doubling in performance relative to the existing logit model (identifying 16 vs. 9 cases), and now they claim a modest improvement in the erratum for that same period (identifying 10 cases vs. 8 cases). When a user runs their updated code, there is not even a modest improvement (identifying 8 cases vs. 8 cases).

Another thing that bothered me

Later, something else bothered me: If an erratum were in order, why didn't they say that in their March 2021 reply to me? In that reply, they act as though nothing in their 2020 JAR article is amiss. Furthermore, when invited by *Econ Journal Watch* to supply a second reply, why didn't they take that opportunity to say that an erratum was in order?

Next, I'll describe what the "error" was, and my journey writing these critiques (the present one being the third). I believe the case supports a conclusion of research misconduct. Notwithstanding the issues I identified in their erratum, the real issue is: *what was this "error" and how did it arise in the first place?*

What was the "error"?

The authors do not make clear to the reader what they are correcting in the erratum. Their erratum starts "Walker (2021a and 2021b) identified an error in the program codes of Bao et al. (2020) posted on Github that led to an overstatement of model performance metrics" and that this error "resulted in approximately 10% of these spanning serial fraud cases in the training period not being recoded as zero." Let me recast this statistic. First, there were only 17 cases affected by this "error" out of 435 total fraud cases in their sample. This entire saga is about what happened to the data for those 17 cases.

In fact, "error" is not the correct word to describe the issue. The code worked fine. Rather, it was the manipulation of the dataset itself where 17 unique fraud cases in the dataset received two identifiers, a manipulation of the underlying data which made no sense.

A journey into criticism

For my dissertation work, I wanted to explore how machine learning might better answer accounting research questions, including fraud detection. I quickly discovered that detection models performed very poorly, but were characterized by the literature as contributing meaningful results that could be useful to industry professionals—which, by the way, was in contradiction to what I have heard anecdotally from those same professionals. As part of my literature review, Bao et al. (2020) was included, but I originally did not pay much attention to it because the original (overstated) results were similar to what I achieved in my own machine learning approach, and those results were still quite poor. As I was wrapping up a draft version of my paper for my committee in late 2020, I decided that, for completeness I should download the code from their paper and try to replicate their results.

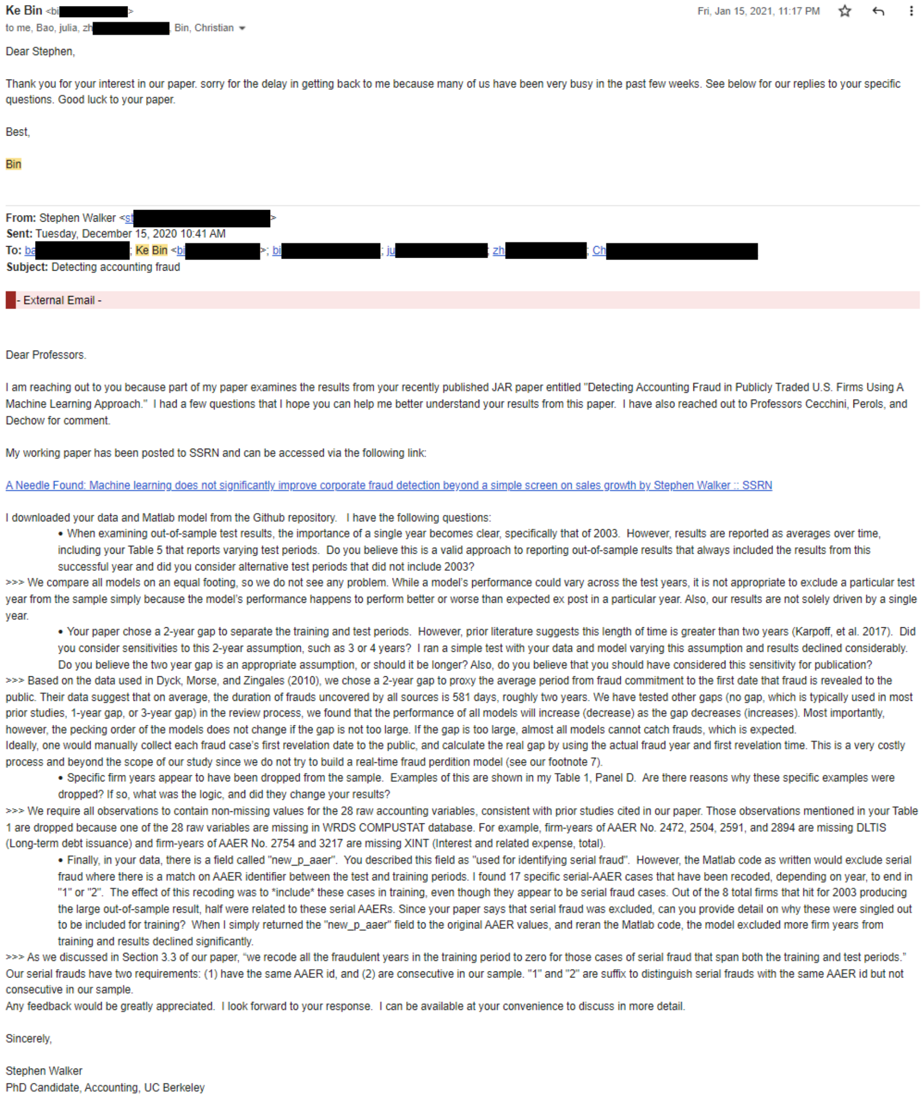
Upon inspection of this code, I noticed a strange new field called *new_p_aaer*, which appeared to identify the fraud cases. I thought *Why would there need to be a new identifier?* The fraud cases already had one. I then downloaded the data into a spreadsheet and pivoted it to see which years and firms were affected by this new identifier. When I saw that it only affected 17 cases, I knew something was amiss. The Bao et al. (2020) article was published in what I believed was one of the top journals in accounting research, the *Journal of Accounting Research*, and the publication included the words "Accepted by Christian Leuz," who was a legend in accounting research at the University of Chicago. At first, I thought there must have been an error in my own thinking. Surely there was a logical explanation. However, as I continued to iterate the problem, I gained confidence that there was a problem here, and it looked like it was done deliberately. However, the authors made no mention in their written paper of why they implemented the *new_p_aaer* manipulation— or even that they did. I first wrote about the issue in a paper that I uploaded to SSRN on November 29, 2020.

In talking around my department about the issue, I was given the name of Alex Young, who had previously written in EJW. His paper criticizing authors Andrew Bird and Stephen Karolyi subsequently led to a retraction at *The Accounting Review*. I first reached out to EJW on December 14. The editor suggested I first reach out to the authors to get their response before writing a journal article. On December 15, I emailed the author group a list of my questions copying their editor, Professor Christian Leuz (Figure 1). Over the next few weeks, I received no acknowledgment or reply for my questions. On January 5, 2021, I received an email from Professor Leuz asking if I had heard back from the authors. I replied no, and then he replied on January 7 that he would "nudge" the authors to answer my questions. On January 15, one month after my initial inquiry, I received the reply from the authors (Figure 1). In my original email, the first few questions were qualitative, about the robustness in the results, but the final question was the most important. The question and reply are below.

Question (email from Walker):

> Finally, in your data, there is a field called "new_p_aaer." You described this field as "used for identifying serial fraud". However, the Matlab code as written would exclude serial fraud where there is a match on AAER identifier between the test and training periods. **I found 17 specific serial-AAER cases that have been recoded, depending on year, to end in "1" or "2". The**

Figure 1. The author group's replies (in entirety), indicated by >>>>, interspersed between the bulleted points of Walker's original email



Ke Bin <bi[redacted]>
to me, Bao, julia, zh[redacted], Bin, Christian

Fri, Jan 15, 2021, 11:17 PM

Dear Stephen,

Thank you for your interest in our paper. sorry for the delay in getting back to me because many of us have been very busy in the past few weeks. See below for our replies to your specific questions. Good luck to your paper.

Best,

Bin

From: Stephen Walker <st[redacted]>
Sent: Tuesday, December 15, 2020 10:41 AM
To: ba[redacted]; Ke Bin <bi[redacted]>; bi[redacted]; ju[redacted]; zh[redacted]; Ch[redacted]
Subject: Detecting accounting fraud

- External Email -

Dear Professors.

I am reaching out to you because part of my paper examines the results from your recently published JAR paper entitled "Detecting Accounting Fraud in Publicly Traded U.S. Firms Using A Machine Learning Approach." I had a few questions that I hope you can help me better understand your results from this paper. I have also reached out to Professors Cecchini, Perols, and Dechow for comment.

My working paper has been posted to SSRN and can be accessed via the following link:

A Needle Found: Machine learning does not significantly improve corporate fraud detection beyond a simple screen on sales growth by Stephen Walker :: SSRN

I downloaded your data and Matlab model from the Github repository. I have the following questions:

- When examining out-of-sample test results, the importance of a single year becomes clear, specifically that of 2003. However, results are reported as averages over time, including your Table 5 that reports varying test periods. Do you believe this is a valid approach to reporting out-of-sample results that always included the results from this successful year and did you consider alternative test periods that did not include 2003?

>>> We compare all models on an equal footing, so we do not see any problem. While a model's performance could vary across the test years, it is not appropriate to exclude a particular test year from the sample simply because the model's performance happens to perform better or worse than expected ex post in a particular year. Also, our results are not solely driven by a single year.

- Your paper chose a 2-year gap to separate the training and test periods. However, prior literature suggests this length of time is greater than two years (Karpoff, et al. 2017). Did you consider sensitivities to this 2-year assumption, such as 3 or 4 years? I ran a simple test with your data and model varying this assumption and results declined considerably. Do you believe the two year gap is an appropriate assumption, or should it be longer? Also, do you believe that you should have considered this sensitivity for publication?

>>> Based on the data used in Dyck, Morse, and Zingales (2010), we chose a 2-year gap to proxy the average period from fraud commitment to the first date that fraud is revealed to the public. Their data suggest that on average, the duration of frauds uncovered by all sources is 581 days, roughly two years. We have tested other gaps (no gap, which is typically used in most prior studies, 1-year gap, or 3-year gap) in the review process, we found that the performance of all models will increase (decrease) as the gap decreases (increases). Most importantly, however, the pecking order of the models does not change if the gap is not too large. If the gap is too large, almost all models cannot catch frauds, which is expected.
Ideally, one would manually collect each fraud case's first revelation date to the public, and calculate the real gap by using the actual fraud year and first revelation time. This is a very costly process and beyond the scope of our study since we do not try to build a real-time fraud perdition model (see our footnote 7).

- Specific firm years appear to have been dropped from the sample. Examples of this are shown in my Table 1, Panel D. Are there reasons why these specific examples were dropped? If so, what was the logic, and did they change your results?

>>> We require all observations to contain non-missing values for the 28 raw accounting variables, consistent with prior studies cited in our paper. Those observations mentioned in your Table 1 are dropped because one of the 28 raw variables are missing in WRDS COMPUSTAT database. For example, firm-years of AAER No. 2472, 2504, 2591, and 2894 are missing DLTIS (Long-term debt issuance) and firm-years of AAER No. 2754 and 3217 are missing XINT (Interest and related expense, total).

- Finally, in your data, there is a field called "new_p_aaer". You described this field as "used for identifying serial fraud". However, the Matlab code as written would exclude serial fraud where there is a match on AAER identifier between the test and training periods. I found 17 specific serial-AAER cases that have been recoded, depending on year, to end in "1" or "2". The effect of this recoding was to *include* these cases in training, even though they appear to be serial fraud cases. Out of the 8 total firms that hit for 2003 producing the large out-of-sample result, half were related to these serial AAERs. Since your paper says that serial fraud was excluded, can you provide detail on why these were singled out to be included for training? When I simply returned the "new_p_aaer" field to the original AAER values, and reran the Matlab code, the model excluded more firm years from training and results declined significantly.

>>> As we discussed in Section 3.3 of our paper, "we recode all the fraudulent years in the training period to zero for those cases of serial fraud that span both the training and test periods." Our serial frauds have two requirements: (1) have the same AAER id, and (2) are consecutive in our sample. "1" and "2" are suffix to distinguish serial frauds with the same AAER id but not consecutive in our sample.
Any feedback would be greatly appreciated. I look forward to your response. I can be available at your convenience to discuss in more detail.

Sincerely,

Stephen Walker
PhD Candidate, Accounting, UC Berkeley

effect of this recoding was to *include* these cases in training, even though they appear to be *serial* fraud cases. Out of the 8 total firms that hit for 2003 producing the large out-of-sample result, half were related to these serial AAERs. Since your paper says that serial fraud was excluded, can you provide detail on why these were singled out to be included for training? When I simply returned the "new_p_aaer" field to the original AAER values, and

> reran the Matlab code, the model excluded more firm years from training and results declined significantly. (boldface and italics added)
>
> **Answer (email from Bin Ke):**
>
> As we discussed in Section 3.3 of our paper, "we recode all the fraudulent years in the training period to zero for those cases of serial fraud that span both the training and test periods." Our serial frauds have two requirements: (1) have the same AAER id, and **(2) are consecutive in our sample. "1" and "2" are suffix to distinguish serial frauds with the same AAER id but not consecutive in our sample.** (boldface added)

That is really a non-answer. It is problematic for the following reasons:

1. The authors disclose an additional rule for identifying serial fraud cases—so as to remove them from training. That additional rule is shown in boldface and was not disclosed in the original publication (and, as I'll show, in violation of the journal's data policy).
2. They completely ignored the effect I described in the question, shown in boldface—that 17 specific cases were included in training, which was in direct contradiction to what their paper described.
3. That now-disclosed rule makes no sense, except to artificially improve the results published.

Their email annoyed me. I proceeded with my initial EJW critique (Walker 2021a). Next, I explain why I believe research misconduct occurred. But, first, I should define research misconduct.

Defining misconduct

A Google search for "academic research misconduct definition" returns a variety of sources, each with varying definitions. Here, I simply turn to what is defined in the law. According to the United States Code of Federal Regulations 42 CFR 93.103 (link):

> § 93.103 Research misconduct.
>
> Research misconduct means fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results.
>
> (a) Fabrication is making up data or results and recording or reporting them.
>
> (b) Falsification is manipulating research materials, equipment, or

> processes, or changing or omitting data or results such that the research is not accurately represented in the research record.
>
> (c) Plagiarism is the appropriation of another person's ideas, processes, results, or words without giving appropriate credit.
>
> (d) Research misconduct does not include honest error or differences of opinion.

I believe that there is substantial evidence that the work in Bao et al. (2020) falls under (b), falsification. Falsification is the manipulation of research materials, in this case, the underlying dataset.

The *Journal of Accounting Research* data policy

The authors are in clear violation of the data policy at the journal (**link**). According to the data policy updated November 2016,

> To be provided upon acceptance of the paper and prior to publication:
>
> 5. **The computer programs or code used to convert the raw data into the final dataset used in the analysis plus a brief description that enables other researchers to use this program.** The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the raw data were processed, the final sample was formed, variables were defined, outliers were treated, etc. This code or programming is in most circumstances not proprietary. However, we recognize that some parts of the code or data generation process may be proprietary, including from the authors' perspective. Therefore, instead of the code or program, researchers can provide a detailed step-by-step description of the code or the relevant parts of the code such that it enables other researchers to arrive at the same final dataset used in the analysis. In such cases, the authors should inform the editors upon initial submission, so that the editors can consider an exemption from the code sharing requirement. Whenever feasible, authors should also provide the identifiers (e.g., CIK, CUSIP) for their final sample. (boldface added)

According to this data policy, unless the editors exempt them for proprietary reasons, they are required to produce the code. In the original zipfile, the authors included three relevant files: The (1) RUSBoost code in Matlab called run_RUSBoost28.m; (2) the SAS code to prepare the final dataset that included the 28 raw accounting variables and the 14 financial ratios called SAS coding.pdf; and (3) BKLYZ Datasheet which described the data. In the BKLYZ Datasheet, the only reference to the *new_p_aaer* field was:

> The variable new_p_aaer is used for identifying serial frauds as described in Section 3.3 (see the code in "RUSBoost28.m" for more details)

As I previously wrote, the authors did not disclose how or why the *new_p_aaer* variable was created. Neither the RUSBoost28.m file nor Section 3.3 of their paper describes it. The SAS coding.pdf file which provided the code to build the dataset (and would be the obvious location for the code to create the *new_p_aaer* variable) contains no reference to it. Clearly, the authors did not disclose the steps required to create this field and hid it from the journal and its readers, in violation of the data policy.

Interestingly, the data policy in effect was changed after the authors' submission on October 7, 2015. Based on an online copy of the older document (**link**) that I found using a Google search, the previous language for the policy dated December 2014 was:

> 5. Prior to final acceptance of the paper, the computer program used to convert the raw data into the dataset used in the analysis plus a brief description that enables other researchers to use this program. ***Instead of the program*, researchers can provide a detailed step-by-step description that enables other researchers to arrive at the same dataset used in the analysis.** The purpose of this requirement is to facilitate replication and to help other researchers understand in detail how the sample was formed, including the treatment of outliers, Winsorization, truncation, etc. This programming is in most circumstances not proprietary. However, we recognize that some parts of the data generation process may indeed be proprietary or otherwise cannot be made publicly available. In such cases, the authors should inform the editors upon submission, so that the editors can consider an exemption from this requirement. (boldface added)

The policy at the time of their submission would have allowed the authors to opt for a step-by-step description instead of the code. However, by the time it was published, the code was required for publication.

Why did the authors create new identifiers?

We still do not know. They have never addressed why creating new identifiers was logical. Even with their erratum, it was hand-waved away as a coding error.

The answer, I believe, has to do with the timeline of publication. The Bao et al. (2020) paper was submitted to the *Journal of Accounting Research* on October 7, 2015, and accepted four years later on October 1, 2019. A clue from this paper tells us that this relabeling of identifiers was not part of the original submission.

Note 10 of the paper says, "We thank the anonymous referee for raising this point." This note is in reference to the preceding paragraph in Section 3.3 "Serial Fraud," which describes the issue: "Such serial fraud may overstate the performance of the ensemble learning method if instances of fraudulent reporting span both the training and test periods." The authors describe their supposed remedy: "we recode all the fraudulent years in the training period to zero for those cases of serial fraud that span both the training and test periods" (Bao et al. 2020, 203).

So we know that the original submission did not correct for serial fraud, and its results were likely substantially better than those published. In my experience using machine learning, it is easy to fool oneself into believing that these models generate substantial improvements over older methodologies. Most of the time there is a data leak that causes overfitting to a test set. So, an anonymous referee pointed this serial fraud issue out to the authors, and they would obviously need to correct it in the next submission. The process to correct for these cases is simple: find where the fraud identifier in the test sample matches in the training sample and recode the indicator variable from one to zero for training purposes. As I showed in my first critique (Walker 2021a), doing exactly that produces null results relative to benchmark models.

So, the authors took a further step, an undisclosed step. Again from their email response:

> Our serial frauds have two requirements: (1) have the same AAER id, **and (2) are consecutive in our sample. "1" and "2" are suffix to distinguish serial frauds with the same AAER id but not consecutive in our sample.** (boldface added)

The result of the requirement in boldface was to usher the 17 serial fraud cases into the training sample—a problem that I made clear in my original email. The effect of including those 17 cases was to boost the number of hits from 9 to the 16 claimed in their paper (see Walker 2021a, Table 2).

In conclusion, evidence of falsification and an ensuing coverup rests on:

1. Creating an absurd and undisclosed requirement, requiring the creation of new IDs to allow serial fraud cases in the training sample that generated misleading and false results, in contradiction to their written paper, and in violation of JAR's data policy.
2. Writing a misleading reply to my initial inquiry of December 15, 2020, where they ignored the problem entirely, even though I clearly described the issue in detail in the email. Furthermore, that inquiry was initially ignored, only to be replied to a month later after the "nudge" by the editor.

3. Writing a misleading reply (Bao et al. 2021) to my critique in EJW where they did not address or attempt to explain the problem, instead bringing up non-germane issues such as the recalibration of trees (which the authors did not previously describe in writing or provide code for, and which, as I show here with their updated code, does not matter).
4. Writing an erratum where they admit to an "error," but mischaracterize it as a coding error. They do not explain what the problem truly was (a direct change to the dataset itself), nor do they describe how they came to this course of action in the first place.
5. Making an obvious false statement describing their original paper in the erratum where they redefine their primary sample period from 2003–2008 to 2003–2005 in order to present new results in the most favorable light possible.
6. Publishing updated code that does not generate the results reported in the tables in the erratum.
7. Publishing that erratum after having had two chances to say in EJW what they say there. (The erratum was posted by JAR approximately 16 months after their March 2021 EJW reply to me.)

For these reasons, on August 10, 2022, I requested an independent investigation to the editors at the *Journal of Accounting Research*. As of September 10, 2022, I have only received an acknowledgement of receipt by JAR editor Leuz who has promised that the editorial board would respond to this request.

Gelman's ladder

Andrew Gelman (2019) of the Columbia Statistics Department describes a ladder of response to sound criticism in academic research. I reproduce this ladder here for discussion purposes. Virtue is high, vice is low.

1. Look into the issue and, if you find there really was an error, fix it publicly and thank the person who told you about it.
2. Look into the issue and, if you find there really was an error, quietly fix it without acknowledging you've ever made a mistake.
3. Look into the issue and, if you find there really was an error, don't ever acknowledge or fix it, but be careful to avoid this error in your future work.
4. Avoid looking into the question, ignore the possible error, act as if it

had never happened, and keep making the same mistake over and over.

5. If forced to acknowledge the potential error, actively minimize its importance, perhaps throwing in an "everybody does it" defense.
6. Attempt to patch the error by misrepresenting what you've written, introducing additional errors in an attempt to protect your original claim.
7. Attack the messenger: attempt to smear the people who pointed out the error in your work, lie about them, and enlist your friends in the attack. (Gelman 2019)

I score the response from Professor Bin Ke to my initial email inquiry and the authors' EJW reply a '4.' They ignored the issue entirely. I score the erratum a '6.' It is a clear attempt to patch the error with misrepresentation of the primary test period.

Journal of Accounting Research editor Christian Leuz on academic misconduct

Not long ago, Professors Luzi Hail of the Wharton School, Mark Lang of the University of North Carolina, and Christian Leuz at the University of Chicago collaborated to write the article "Reproducibility in Accounting Research: Views of the Research Community" (2020). They write in response to a 2019 survey conducted at a *Journal of Accounting Research* conference: "Most have encountered irreproducibility in the work of others (although not in their own work) but chose not to pursue their failed reproduction attempts to publication" (Hail, Lang, and Leuz 2020, abs.). They consider three paths forward:

1. Business as usual.
2. Public repositories of code and data.
3. Increased incentives for replication.

In the discussion under "business as usual", Hail, Lang, and Leuz indicate that "a substantial minority of respondents" favor business as usual, as "the status quo reflects market forces, weighing the costs and benefits for authors, journals, and universities" (2020, 17).

As for the second path forward of having a repository for code and data, I am thankful for the code and data being available, which allowed me to write the critique. For the third remedy, increased incentives, they write, "As example for a sub-component in an existing outlet, the *Journal of Finance* (JF) includes a section for “Replications and Corrigenda” that provides space for “short papers that document material sensitivities of central results in papers published in the JF.” If the *Journal of Accounting Research* had this remedy, I likely would have submitted my critiques there. However, as of now, they do not have a credible and independent process for submitting criticism of their articles. My hope is that a future editorial board would embrace criticism—it is vital to scientific credibility.

Appendix

Data and code related to this research is available from the journal website (**link**).

References

Bao, Yang, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. 2020. Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research* 58(1): 199–235.

Bao, Yang, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. 2021. A Response to “Critique of an Article on Machine Learning in the Detection of Accounting Fraud.” *Econ Journal Watch* 18(1): 71–78. **Link**

Bao, Yang, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. 2022. Erratum. *Journal of Accounting Research* 60(4): 1635–1646. **Link**

Gelman, Andrew. 2019. A Ladder of Responses to Criticism, from the Most Responsible to the Most Destructive. *Statistical Modeling, Causal Inference, and Social Science*, January 18. **Link**

Hail, Luzi, Mark Lang, and Christian Leuz. 2020. Reproducibility in Accounting Research: Views of the Research Community. *Journal of Accounting Research* 58(2): 519–543.

Walker, Stephen. 2021a. Critique of an Article on Machine Learning in the Detection of Accounting Fraud. *Econ Journal Watch* 18(1): 61–70. **Link**

Walker, Stephen. 2021b. Rejoinder to the Critique of an Article on Machine Learning in the Detection of Accounting Fraud. *Econ Journal Watch* 18(2): 230–234. **Link**

About the Author

Stephen Walker earned his Ph.D. at the University of California Haas School of Business in May 2021. Prior to his Ph.D. studies, Stephen worked in equity research at Sanford C. Bernstein in New York City. He also holds an MBA from Columbia Business School. He currently works as an economic consultant for shareholder litigation and can be reached via his personal website at stephenwalker.me.

Go to archive of Comments section
Go to September 2022 issue

Discuss this article at Journaltalk:
https://journaltalk.net/articles/6054/