Goldfarb, B., & Yan, L. (2021). Revisiting Zuckerman's (1999) categorical imperative: An application of epistemic maps for replication. Strategic Management Journal, 42(11), 1963-1992. https://doi.org/10.1002/smj.3290

Revisiting Zuckerman's (1999) Categorical Imperative: An Application of Epistemic Maps for Replication

Brent Goldfarb*
Associate Professor of Management and Entrepreneurship at the University of Maryland's Robert H. Smith School of Business;
Academic Director at the Dingman Center for Entrepreneurship
University of Maryland
7621 Mowatt Ln
College Park, Maryland 20742

Liyue Yan
Robert H. Smith School of Business
University of Maryland
7621 Mowatt Ln
College Park, Maryland 20742
College Park

January 2021

Running Head: **Revisiting Zuckerman's (1999) Categorical Imperative**

Keywords: Categories, Legitimacy, Replication, Epistemic Maps, Model Uncertainty

*Corresponding author. For inquiries, write to brentg@umd.edu. We thank Ezra Zuckerman for helpful comments and graciously sharing data in support of this project. We thank conference participants at Strategic Management Society Annual meetings in Houston (2017), seminar participants at Aalto University, and helpful comments from colleagues at the Maryland Student Presentation Seminar. Financial support of the Richard M. Schulze Foundation was appreciated. All errors are the authors'. Code and procedures used for downloading subscription data sets can be found here: https://osf.io/b6qsy/?view_only=65a193b736e544bcb21cdf5a6a6c9c21

ABSTRACT

Academic Abstract*:* We revisit Zuckerman's (1999) "*The Categorical Imperative: Securities Analysts and the Illegitimacy Discount*," which theorizes that when organizations are recognized as legitimate players in a category, they perform better. A replication exercise fails to reproduce two of three sets of results. Assisted by data shared by the original author, we find evidence that the inconsistency is due to a coding error in the original and differences between analyst data sets. We illustrate the use of epistemic maps and evaluate the theory's predictive power across a broad set of plausible empirical assumptions and also for a subsequent time period. The results are not robust. We conclude that these data provide little evidence to support strategic recommendations. Challenges and remedies for replication are discussed.

Managerial Abstract: This paper replicates a 1999 study "*The Categorical Imperative: Securities Analysts and the Illegitimacy Discount"* (Zuckerman, 1999). A key implication of the 1999 study is that firms should actively seek to be covered by analysts who specialize in their industries, and that a failure to be covered by analysts who are industry specialists will lead to lower capital market valuations. Our replication exercise indicates that there is insufficient evidence to recommend actively managing coverage along this dimension.

1. INTRODUCTION

Zuckerman's (1999, hereafter Z99) seminal study "*The Categorical Imperative: Securities Analysts and the Illegitimacy Discount"* theorizes that legitimacy is bestowed upon properly categorized actors. Z99 examines whether firms who are improperly categorized, in that they fail to attract securities analysts that are industry specialists, will be undervalued in capital markets.

This theoretical framework of categories has had a significant influence on strategy research. As of July 2020, Z99 has been cited over 2,000 times, with 1,000 of these citations occurring since 2014. Z99 is influential because it goes beyond the simple supply-demand lens used to understand the interplay among market participants. Instead, it argues that an organization is first not evaluated by the features of its products, but through the space, or category, in which it claims to be. Since non-recognition by a category's audience is hypothesized to delegitimize and devalue an organization, how firms are categorized may critically influence their performance and survival.

The theory thus implies that how a firm is categorized is an important strategic concern for a firm. Indeed, since Z99, many studies have examined category-related questions with strategic implications including: What is the influence of categorization or identity on audience understanding and evaluation (e.g., Hsu, 2006; Durand, Rao, and Monin, 2007; Smith, 2011; Pontikes, 2012; Negro and Leung, 2013; Wry, Lounsbury, and Jennings, 2014; Lo and Kennedy, 2015; Paolella and Durand, 2016)? How do analysts mediate audience evaluation and how do firms react to this (e.g., Zuckerman, 2000; Benner and Ranganathan, 2012; Feldman, 2016)? How do firms choose market categories, de-diversify, build entrepreneurial identities and construct new market categories (e.g., Zuckerman, 2000; Navis and Glynn, 2010; Khaire and Wadhwani, 2010; Navis and Glynn, 2011; Litov, Moreton, and Zenger, 2012; Granqvist, Grodal, and Woolley, 2013)?[1]

We first replicate the analysis in Z99. Under some assumptions, we find estimates consistent with a categorical discount, though the measured effects sizes are much smaller than those reported in Z99. Under other assumptions we do not. A systematic forensics analysis leads

[1] Vergne and Wry (2014) identify Zuckerman's contribution to the categorization literature as seminal because it was the first to view categories as the carriers of the cultural infrastructure that audiences use to understand and evaluate organizations. They note that Zuckerman's work has also given rise to a sociological view of industry as a category of organizations that share common features but at the same time differentiate to compete.

us to believe that our differences are due to a combination of an error made in the original paper in the dependent variable calculation, and differences between the Zacks and I/B/E/S analyst databases. We then employ an epistemic mapping approach that links empirical assumptions to estimates (King, Goldfarb, and Simcoe, 2020, Simonsohn, Simmons, and Nelson, 2015; Sala-i-Martin, 1997; Leamer, 1983). The approach attenuates a central problem in the epistemology of testimony on empirical reports which Gelman and Loken (2013) dub the "garden of forking paths." Each "path" refers to a specific set of assumptions including empirical sampling scheme, specification, estimation technique, and mapping between constructs and measures. Through this we identify which combinations of assumptions are required to find a relationship consistent with the categorical discount theory and, conversely, which combinations of assumptions lead to estimates less consistent with the theory.[2]

Our intention is not to challenge Z99's theoretical ideas. Rather, we carry out an exercise to understand whether the accompanying empirical results reported in Z99 are sensitive to alternative judgments other researchers might have made. With this, we can evaluate to what extent the results in Z99 should be viewed as a basis for creating cumulative knowledge (Bettis, Ethiraj, Gambardella, Helfat, and Mitchell, 2016), or as a basis for strategic recommendations. Although one of our data sources differs, our replication of the specifications reported in Z99 is best thought of as a narrow replication in that we draw from the same population, and our samples are mostly overlapping (Bettis, Helfat and Shaver, 2016).

We quickly review Z99's theoretical proposition. Z99 proposed: "*Ceteris paribus, a product experiences weaker demand to the extent that it does not attract reviews from the critics who specialize in the category in which it is marketed" (p. 1405).* The proposition was

[2] The forking paths problem is a problem of testimony because it is difficult to convey the judgments that lead a researcher to choose some paths over others, and because judgments, by definition, are subject to challenge (Longino, 1990, Wilholt, 2013). Agreement on the most appropriate model is important because interpretation of reported frequentist statistics requires the assumption that the estimated model is true (Spanos, 2014). If a reader and a researcher disagree about whether a model is appropriate, they will also disagree on its interpretation. However, if a reader cannot ascertain the multitude of assumptions and judgments that led an author to report a small subset of models, they will be unable to evaluate whether they would or would not disagree. Moreover, with only a narrow reporting of specifications, the reader will be unable to evaluate the sensitivity of the conclusions to researcher's judgments. In contrast, epistemic maps of a large number of plausible models, sampling and measurement assumptions to empirical results, allow the reader to evaluate robustness across many potential judgements. Maps thus enable the reader to evaluate the implications of subsets of the assumptions that they may find most plausible. Epistemic mapping therefore shifts some of the inferential work from the researcher to the reader as opposed to compelling the reader to accept or reject a small set of presented models. Our operationalization of this approach draws from Berchicci and King (2020).

operationalized in the context of security analysts and a hypothesis was tested: "*Ceteris paribus, the greater the coverage mismatch suffered by a firm, the lower its stock price" (p.1412).*[3] We reconstruct Z99's data set from the original data sources and then, as in Z99, measure the statistical relationship between measures of excess value and Z99's proposed measure of coverage mismatch. Due to data availability, in our replication we collected analyst data from the Institutional Brokers' Estimate System (I/B/E/S) as opposed to Z99's original source (Zacks). Both data sets have been used frequently in research in financial economics.

We then summarize a series of replications of the specifications chosen in Z99. We present the results for three data processing procedures because there is some ambiguity as to what was implemented in Z99. Some estimates are consistent with the central hypothesis, while others are not. A subset of the estimates is also consistent with those reported in Z99 in terms of magnitude. With the original author's help, we were able to compare our measures to Z99's for a small subsample. We reason that these differences are likely a combination of differences in an underlying data source, as well as an error in the dependent variable construction in Z99.

We have a greater interest in whether the theory robustly predicts patterns in the data, as opposed to whether we can replicate the specific models in Z99. We produce a series of epistemic maps of 10,361 models that relate excess value to coverage mismatch for firms in Z99's study period, 1985–1994. Each regression links a set of empirical assumptions regarding sample construction, measurement of constructs and choice of controls to an estimated coefficient relating coverage mismatch to firm value. The maps demonstrate that the results are sensitive to these empirical assumptions. The combinations of controls that produce the estimates most consistent with the theory for one dependent variable and sample construction choice, generally do not produce estimates similarly consistent with an alternative dependent variable and sample construction choice. While many point estimates are consistent with the theory in terms of sign, few estimates match the magnitude of the 12 reported Z99 estimates.

We then repeat the exercise with a later sample, 1995–2004. As in the 1985–1994 map, some point estimates are negative but small relative to those reported in Z99. And again, for this later sample, the results are also sensitive to the choice of sampling strategy, the measurement of

[3] Empirically, Z99 does not identify a causal linkage—which would have required taking into account the analysts' choices to cover firms. Instead, the study investigates market valuations conditional on such choices. We do not solve this endogeneity problem in this paper, though this is an additional barrier to linking Zuckerman (1999)'s empirical exercise to the theory.

coverage mismatch, and controls. We also find that the sets of assumptions which produce the estimates most consistent with the theory the 1985–1994 period do not do so in the later period 1995–2004, and vice versa. We interpret the findings as follows: For a researcher to conclude that the data are largely consistent with the theory, they must believe either that the most appropriate methods to measure constructs and the most appropriate controls to include in the estimations are different across dependent variables and periods, or that some dependent variables and periods should be discounted altogether.[4]

We cannot share our data due to data licensing restrictions. However, they are available at many research institutions. We publish our code and the epistemic maps. An interested reader will be able to select the subset of models they find plausible and easily understand the distribution of associated estimates in this subset.

We draw three central conclusions. First, there is insufficient evidence in these data to make recommendations to CEOs or other managers that they should try to manage whether their firms are covered by particular analysts. The inconsistency in the results across different assumptions suggests that the original results are largely due to sampling variation. Second, more extensive reporting, including sharing of code and data, will help mitigate the problem of testimony. When data cannot be shared publicly, we urge authors to maintain copies of their data and code to facilitate later replication. Third, authors can mitigate the necessity of replication by reporting a broader analysis of robustness of results across different assumptions. This can clarify the sensitivity of results and allow readers to draw conclusions if they prefer different assumptions than those originally (or initially) published in a paper. To proceed, in Section 2, we replicate and report results corresponding to Z99's empirical exercise. In Section 3, we report on the comprehensive mapping of assumptions to results. In Section 4, we reflect on implications for replication. In Section 5 we conclude.

2. REPLICATION

2.1 Data and sample

Three data sets are used to replicate Z99's empirical exercise: The Center for Research in Security Prices (CRSP), Compustat, and I/B/E/S. The CRSP data provide daily stock market information, including stock prices and shares outstanding. Compustat provides financial

---

[4] We note that discounting the later period may be reasonable because the calculation of coverage mismatch may have been impacted by changes in SEC-mandated reporting requirements.

information on firms and their business segments such as sales, assets, operational profits, etc. Since the Zacks data used in Z99 are not available to us, we use I/B/E/S data. As described in the Appendix A2-3, I/B/E/S appears more complete than Zacks for the time period of our study.

Our replication exercise is not easily classified by the Bettis-Helfat-Shaver terminology of narrow and quasi replications. Our sample overlaps considerably with Z99's, but we use different data to construct the key independent variable. Thus, our exercise has characteristics of a narrow reproduction and those of a different-test-same-sample quasi replication (Bettis et al., 2016). In Section 3 we conduct a narrow replication of our epistemic mapping exercise using a later time period.

Given that not every detailed technique or assumption is described in the original study, we believe the best way to replicate is to conduct analyses with combinations of different techniques and assumptions. As we describe below, both the dependent and independent variables of interest are a function of both the focal firm and other firms in its industry. For this reason, whether sample exclusion criteria are implemented before or after the calculation of these variables will affect their values. For example, Z99 operationalizes the dependent variable "stock market value" by using a measure "excess value" that had been developed in financial economics (Berger and Ofek, 1995). Excess value is a function of a firm's value relative to the median value of a firm in its industry. A researcher might judge that a class of firms (e.g., foreign firms) should be included in the calculation of this median, but not in the final regression sample. Alternatively, one might exclude foreign firms from both the median calculation and the regression sample. To understand if the results are indeed sensitive to such judgments, we present and evaluate the results for three different procedures that vary the exclusion restrictions and when they are implemented. Table 1 provides a summary of the differences of the three sample construction procedures described presently and the five diagnostic procedures we use in Section 2.4.

Table 1 here

Procedure 1 follows the description in Z99 and excludes firms with any segment in the finance industry, non-US firms, and firms not traded on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and NASDAQ (Z99, p. 1422).[5] Firms with sales less than

---

[5] Firms with financial segments were excluded because many such corporations do not report their earnings before interest and taxes (EBIT), which is needed to construct the dependent variables. Berger and Ofek (1995) stated that

$20 million are not used in the calculation of industry medians. However, these small firms are included in the final analysis, because Z99 only specifies that these firms are excluded from the calculation of interim values necessary to calculate the dependent variables (Z99, p. 1415).

Procedure 2 includes firms with any segment in finance, non-US firms, and firms not traded on the three major exchanges and excludes firms with less than $20 million in sales. This procedure is included because it reflects our best guess, based on the shared data files provided by the original author, of the procedure implemented in Z99. In those files, the independent and dependent variables were populated for firms with segments in finance, non-US firms, and firms traded outside the three major exchanges. This suggests that they were only excluded from the regression sample *after* the interim values had been calculated. In addition, Z99 specifies excluding firms with sales less than $20 million in the calculation of median ratios by industry. This criterion does not automatically exclude firms that have sales less than $20 million from the analysis sample—as such, Procedure 1 includes these firms.[6] However, based on the shared files, it appears that firms with sales less than $20 million were excluded in the calculation of the dependent and independent variables in Z99, and hence they are excluded from the sample in Procedure 2.

Procedure 2a includes firms with financial segments in the regression sample. As noted in Z99 as well as Berger and Ofek (1995), firms with segments in finance were dropped before the calculation of excess value because their reporting of earnings before taxes and interest (EBIT) differs. However, dropping firms with segments in finance may influence the results simply because we are excluding a set of large firms. To evaluate whether this is the case, Procedure 2a modifies Procedure 2 and includes firms with financial segments in the final regression analysis. Note that including firms with financial segments can only be implemented for Procedure 2, as the dependent and independent variables are not calculated for these firms in Procedure 1.

2.2 Variables

2.2.1 Dependent Variable: Excess Value

firms were not required to report segment level data for subsidiaries in the utilities industries. However, we do observe segments in the utilities industry (SIC code 490–499) in our current data set. Our current analyses include these firms. Excluding or including these firms do not affect our results.

[6] Hence, we include these smaller firms in *Procedure 1*.

As in Z99, we follow Berger and Ofek (1995) to calculate "excess value." Excess value is the discrepancy between the estimated value of a firm and its true value. This idea has been operationalized in three primary ways: excess value based on sales, excess value based on assets, and excess value based on earnings. Specifically, the excess value of a firm is measured as the log of the ratio of its total capital V to its estimated value $I(V)$: $ln\ (\frac{V}{I(V)})$. The total capital of a firm is calculated as the sum of the market value of a firm's common stock—the number of shares outstanding multiplied by the share price at its fiscal year's end—and the book value of its debt, denoted as V.[7]

A firm's estimated value $I(V)$ is calculated as the summed estimated value of all its segments, in the function below:

$$I(V) = \sum_{i=1}^{n} A_i \times (Ind_i \left(\frac{V}{A_i}\right)_{mf})$$

where A_i denotes a firm's sales (or assets, or Earnings Before Interest and Taxes (EBIT)) in its segment i, and $Ind_i \left(\frac{V}{AI}\right)_{mf}$ is the median ratio of firm total capital to sales (or assets, or EBIT) in segment i's corresponding industry. Median ratios are industry-level variables calculated based on the exclusion criteria described above.

We calculate the ratio of each firm's total capital to its sales (or assets, or EBIT, or Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA)) in an industry, and take the median as the median ratio $Ind \left(\frac{V}{A_i}\right)_{mf}$. The goal is to measure the market valuation, defined as total capital, of each firm relative to its industry median. However, because many firms participate in multiple industries, and their total capital cannot be divided into different segments, as in Z99 and Berger and Ofek (1995), we only consider single-segment firms in calculating the ratios. We use accounting items in a firm's four-digit Standard Industrial Classification (SIC) class. If the industry had more than five single-segment firms and each firm had sales over $20 million, then the median ratios of these single-segment firms were used. If an industry defined by all firms in a four-digit SIC class did not meet the requirements above, then

[7] To calculate the book value of debt, we calculate the sum of long-term debt (DLTT) and short-term debt (debt in current liabilities; DLC).

the industry medians were calculated at the three-digit SIC level. If an industry did not meet these requirements at the three-digit SIC level, the medians were calculated at the two-digit level.

We follow Z99 in making the following adjustments. The summed, segment-level data can be different from the firm-level data due to accounting decisions. For firms that have a discrepancy greater than one percent of sales data, all types of excess value are treated as missing. For firms that have discrepancy greater than 25 percent on assets or EBIT data, excess value is treated as missing on the corresponding variable. For the firms that have discrepancy within 25 percent, segment variables are recalculated by multiplying the segment's proportion of the segment sum for that variable by the firm's total.

Both Z99 and Berger and Ofek (1995) use Earnings Before Interest, Taxes and Depreciation (EBITD) to replace negative values of EBIT. However, because the current Compustat segment data set does not provide accounting numbers to calculate EBITD, we use the next closest measure, EBITDA.[8] As in Berger and Ofek (1995), when the EBIT-based estimation cannot be calculated directly due to a segment's negative EBIT, we calculate the EBITDA-based estimation if EBITDA is positive. If EBITDA is still negative, then for that segment we use the sales-based estimation to replace the EBIT-based estimation. Though EBITD and EBITDA are similar, they are not the same; hence we expect a discrepancy between ours' and Z99's EBIT-based dependent variables.

2.2.2 Independent Variable: "Coverage Mismatch"

The independent variable "coverage mismatch" measures the extent to which a firm is ignored by industry specialists. The theory argues that industry specialists mediate between firms and investors and shape investors' understanding of the stock market. If a firm is not covered by enough industry specialists, it will be less likely to attract attention from investors. The mismatch measure is independent of the total number of analysts covering a particular firm relative to other firms in the industry. Rather, it is a function of the number of specialists covering a firm relative to the total number of specialists in an industry.

An analyst is considered a specialist if they cover a sufficient number of industry firms, where sufficiency is a function of industry size. Following Z99, three steps are involved: (1) A firm is allocated to an industry if it has its highest proportion of its sales in that industry. (2) The

[8] To calculate EBITDA at the segment level, we calculate EBITDA as the sum of operating profit, depreciation, and amortization.

total number of industry-based firms that have ever been covered by an analyst in the industry is then calculated. (3) Analysts are defined as industry specialists if they cover a sufficient number of firms in an industry, where what qualifies as sufficient is a function of the size of the industry (see Table 2 reproduced from Z99 Table 1). Industries are defined with three-digit SIC classes. All specialist criteria are updated annually based on a fiscal calendar.[9]

Coverage mismatch is measured at the firm-segment level. For firm f's segment-level in industry i, coverage mismatch, denoted by cm_{fi}, is:

$$cm_{fi} = 1 - \frac{C_{fi}}{\left(C_{gi}\right)}$$

C_{gi} is the number of industry specialists in industry i covering firm g and $\left(C_{gi}\right)$ is the maximum taken over all firms in industry i. The calculation of $\left(C_{gi}\right)$ was done subject to the "best guess" sample exclusion restrictions described above. Segments without any specialists receive a mismatch score of "1." Firm-level coverage mismatch is the segment-sales weighted average of coverage mismatch:

$$cm_f = \sum_{s=1}^{S} w_{fs} \times cm_{fs}$$

where S is the number of segments reported by firm f and w_{fs} is the proportion of total sales.

Table 2 here

2.2.3 Control Variables

We include the complete set of control variables from Z99: the log of a firm's assets as a measure of firm size, the EBIT-to-sales ratio as an indicator of firm profitability, the capital expenditure-to-sales ratio as a control for firm growth opportunities, the sales-based Herfindahl

[9] Firms often have different fiscal calendars within one industry. As the data are panel in nature, Zuckerman constructed a universal fiscal calendar so he could compare them within a time frame. We follow this methodology. This calendar allows us to compare firms' performance in two close periods, which overlap between one and 12 months. In year *t*, any firm with fiscal year-end before June 1st is in its fiscal year t-1 before its fiscal year-end month, while a firm with fiscal year-end after June 1st is in its fiscal year t before its fiscal year-end month. A firm can set its fiscal year-end at any month during a year. Thus, two firms in the same fiscal year (based on their own fiscal calendars) may overlap between one to 12 months. This "universal fiscal calendar" might lead to measurement error. To see this, assume firm A and B are the only two firms in an industry: firm A has a fiscal year-end in May, while firm B has fiscal year-end in June. In December 1985, firm A was in its fiscal year 1985, while firm B was in fiscal year 1986. Accordingly, if an analyst only covered firm A in December 1985 during 1984–1986, this analyst is considered to have covered one firm in 1985; if another analyst only covered firm B in December 1985 during 1984–1986, that analyst is considered not to have covered any firms in 1985. Therefore, becoming an industry specialist in a particular "year" also depends on the timing of coverage and the fiscal calendar a firm uses. However, unless this causes systematic bias, measurement error will be classic, and would then attenuate any findings. We have no reason to believe this problem is more or less severe in the replication than the original.

index as a measure of firm diversification, the R&D expenditure-to-sales ratio as an additional control to capture firm growth, and the number of analysts who covered firm f in a year as a control for the magnitude of analyst attention. In the mapping exercise in Section 3, we also include year fixed effects as an alternative control.

2.3 Replication

Z99 reports 12 specifications that relate excess value to coverage mismatch. There are Four specifications for each of the three alternative excess value measures: excess value based on sales (Models 3–6), excess value based on assets (Models 9–12) and excess value based on earnings (Models 15–18). These four specifications are as follows:

$$EV_k = \beta_0 + \beta_1 cov.mismatch + \beta_2 Log(Assets) + \beta_3 \frac{EBIT}{Sales} + \beta_4 \frac{Cap\ Exp}{sales} + \beta_5 Herf + \beta_6 \#Analysts + \beta_7 \frac{R\&D}{sales} \quad (1)$$

$$EV_k = \beta_0 + \beta_1 cov.mismatch + \beta_2 Log(Assets) + \beta_3 \frac{EBIT}{Sales} + \beta_4 \frac{Cap\ Exp}{sales} + \beta_5 Herf + \beta_6 \#Analysts \quad (2)$$

$$EV_k = \beta_0 + \beta_1 cov.mismatch + \beta_2 Log(Assets) + \beta_3 \frac{EBIT}{Sales} + \beta_4 \frac{Cap\ Exp}{sales} + \beta_5 Herf + \beta_6 \#Analysts + \beta_8 \frac{EBIT}{Sales(t-1)} \quad (3)$$

$$EV_k = \beta_0 + \beta_1 cov.mismatch + \beta_2 Log(Assets) + \beta_3 \frac{EBIT}{Sales} + \beta_4 \frac{Cap\ Exp}{sales} + \beta_5 Herf + \beta_6 \#Analysts + \beta_9 EV_{t-1} \quad (4)$$

where k is *sales, assets* or *earnings* and the control variables are as described above.

We are interested in two questions. First, are the $\hat{\beta}_1$ replication results consistent with those reported in Z99? Second, are the estimated $\hat{\beta}_1 s$ consistent with the theory?

2.3.1 Replication Consistency

Following the journal's guidelines, we refrain from reporting asterisks and report 95 percent confidence intervals of all estimates. We consider a replicated coefficient consistent and the replication successful when the confidence intervals overlap across estimates. Importantly, since we are comparing the estimates to each other, results can be consistent or inconsistent, regardless of whether we can reject the null hypothesis value of zero in a traditional frequentist test.

A claim of consistency means that, if the model is true, and additional random samples are drawn from the same population, 95 percent of these coefficient estimates are expected to be in the confidence interval. The overlapping confidence interval test is weak, in the sense that it is less likely to reject the null hypothesis that the estimates are the same than a Student's t-test (Schenker and Gentleman, 2001). Thus, the overlapping confidence interval criteria is deferential to the original result. Note that for this frequentist test, the Z99 specifications serve as a pre-registration, and constrain the researcher's degrees of freedom. However, for pre-registration to be successful, there should be zero degrees of freedom. In our case, although the specification itself is binding, there is some ambiguity as to how the sample was constructed. We report three replication procedures to understand if the results are sensitive to the sample construction procedure. If the results are invariant to these procedures, then we can be confident that the replication is successful.

2.3.2 Theoretical Consistency:

A claim of theoretical consistency requires, at minimum, that, $\hat{\beta}_1$ is negative. Traditionally, we would thus conclude if the coefficient was statistically significant at a 95 percent level, or, equivalently, if the confidence interval did not include 0. However, frequentist tests must adhere to strict pre-specification requirements (King et al., 2020). Alternatively, as King et al. argue, if the result is negative for all plausible models and data processing variations, we might also be more confident of a conclusion of theoretical consistency. Additionally, some, including this journal, add a criterion of economically meaningful effects. In the reported results, if the confidence intervals include the number zero, a properly pre-specified frequentist test would fail to reject the null hypothesis that there is no effect at the p=0.05 level.[10]

2.3.3 Exact Replication.

In Table 3 we present summary statistics together with the Z99 results for easy comparison. With the exception of the three dependent variables, our sample is about 10 percent smaller than Z99.[11] The control variables, Log of Assets, EBIT/Sales, Capital Expenditures/Sales, Herfindahl, and R&D/Sales, are close to identical to those reported in Z99. For the purposes of a concise discussion, Table 3 reports summary statistics for Procedures 2 and 3 for the dependent variables. The conclusions of the present discussion are similar if we were to consider the summary statistics of Procedures 1 or 2a. Full summary statistics for all procedures described in Table 1 can be found in Appendix 1. The mean number of analysts is slightly higher in the replicated sample, suggesting broader coverage of analysts in I/B/E/S than Zacks. The means of the dependent variables differ by about 0.1 between Z99 and Procedure 2. These differences in means are small relative to standard errors of these statistics and the standard deviations (SD) reported in Z99 are about 0.3 SD larger for the sales- and assets-based measures,

[10] As described in King et al. (2021), interpretation of frequentist statistics requires pre-specification such that the researcher has zero degrees of freedom. In practice, this implies pre-specifying sampling, measurement, and testing protocols prior to collecting data. It is for this reason that, for example, clinical trials are only viewed as valid if they pre-specify and strictly adhere to their pre-specified protocols.

[11] As described in Section 2.4, Zuckerman shared two recovered files with us. One includes fiscal year 1984. It is possible that Z99 included year 1984 in the sample construction and reported what is the them here? in the summary statistics table. We report a duplicate table including FY1984 as Table A2-4. Including FY1984 has minimal effect on the calculated statistics.

and slightly smaller for the earnings-based measure. On the one hand, the differences in the means of the dependent variables appear small relative to their standard deviations. On the other hand, the inability to closely match the dependent variables, the differences in analyst coverage, while at the same time the close matching the control variables prompted considerable forensics, which are discussed in Section 2.4.

Table 3 Here

We summarize the exact replication results for the coefficient of interest, $\hat{\beta}_1$, as well as the estimates reported in Z99 in Table 4, Panel A. Full regression results for all regressions summarized in the table are found in Appendix 1. We infer two central findings. First, under the assumptions of Procedure 1, we would be unlikely to conclude that there is evidence of a directional or an economically important relationship between coverage mismatch and firm valuation. Most estimates are close to zero, and confidence intervals include 0 for all models.

Table 4 here

For Procedure 2, our point estimates are generally negative but much smaller than Z99, and not all would reject the null hypothesis in a properly pre-specified frequentist test. In Model 3, though the entirety of the 95 percent confidence interval of the coefficient on coverage mismatch is negative, Z99's estimate is almost four times larger in absolute value (-0.15 [-0.22, -0.09] vs. -0.04 [-0.07, -0.01]) and the confidence intervals do not overlap. Thus, our replication fails the overlapping confidence intervals test. In Models M4, M5, and M6, the Z99 estimates are between five and 10 times larger in absolute value. In Model M6, the point estimate is very close to zero, and the confidence interval gives little clue as to the expected sign of a repeat estimate.

Continuing with Procedure 2, using the assets-based dependent variable, in all models the coefficients for coverage mismatch are precisely estimated as negative. However, in absolute value, these coefficients are about 75 percent smaller than Z99. In Models M9–M11, the replicated coefficients fail the overlap test. In Model M12, the confidence intervals just overlap. In particular, the confidence interval of the original estimation was [-0.173, -0.046] whereas in the replication it is [-0.047, -0.007].

Using excess value based on earnings, all of our models successfully replicate the original study (Models M15–M18). In this case, we replicate Z99 quite closely in magnitude, and the estimates overlap across all four models. To summarize, using Procedure 2, we were able to reproduce one of the three sets of results from Z99, and for the other two, the results are consistently negative but much smaller in magnitude. Of the 12 specifications, we would fail to reject the null hypothesis in four if we had conducted a properly pre-specified test. Procedure 2a, which is very similar to Procedure 2, produces an almost identical result.

The estimates based on the sales- and assets-based measures imply much smaller economic effects than those reported in Z99. Z99 illustrates the economic significance of the measured effects using the case of PepsiCo in 1985, whose excess value was 0.42 (total capital was $6.57 billion, and sales-based imputed value was $4.3). Whereas the Z99 estimates imply that PepsiCo's valuation would increase 7.2 percent if its coverage mismatch had been 0 instead of 0.51, the replication estimates suggest this number is one percent.[12] We find a much smaller economic effect than reported in Z99.

2.4 Why are our results different?

We conducted several additional checks to evaluate why there are differences in the independent and dependent variable calculations and whether these differences may have affected our results. To this end, we were significantly aided by files generously shared by Zuckerman. The fact that we are using similar data sources allows focus on three potential causes: (1) the construction of the dependent variables and sample differences, (2) changes in overlapping data sources, in particular, potential updates or retrospective changes to Compustat or CRSP, or improved linkages between them, (3) differences between I/B/E/S and Zacks analyst data.

To evaluate (1), whether there were differences in the calculation of the dependent variables, we conducted a partial replication of Berger and Ofek (1995) with the idea that a successful replication would be an indication that our dependent variable calculation is accurate. The replication was indeed successful, which led us to conclude that our dependent variable

[12] The 0.136 estimated coefficient in Z99, Model M4 implies that if PepsiCo's coverage mismatch had been 0 instead of 0.51 (as reported in Z99), PepsiCo's total capital would have been 7.2% higher or $7.04 billion. In Procedures 2 and 2a we estimate a coefficient of .02, which would imply that PepsiCo's total capital would have been 1% higher, or $6.64 billion ($e^{0.51*0.02} * 6.57B$); the implied total capital difference is 14% that of Z99 ($67M instead of $472M). More generally, and keeping with the specification chosen in Z99 to illustrate the magnitude of the results, a one standard deviation decrease in coverage mismatch *(-0.32)* would represent a 0.6% increase in total capital for the average firm based on the point estimate in Procedure 2, Model M2 ($e^{0.32*0.02}$), a 1.1% increase in total value based on the point estimate in Model M10, and a 2.4% increase based on the point estimate in Model M16. We calculate the average effect. For example, applying a one standard deviation decrease to the coverage mismatch requires the initial mismatch value to be at least 0.32, since mismatch is bounded from below by 0. Hence the interpretations discussed here condition on coverage mismatch being above 0.32.

calculations are accurate. We refer the reader to Appendix A2-1 for a detailed report of this exercise.

To address (2), we researched whether there were retrospective data coding changes in Compustat, CRSP or the linkages between the two. We were able to evaluate these possibilities with the aid of the Zuckerman shared files and help from the Standard & Poor's technical support team. For example, if firm industry classifications were retrospectively changed, this would undermine the calculations of the independent and dependent variables. In the Appendix A2-2, we describe our investigation, which reveals no evidence that classifications of firms were retrospectively changed over time. We do find some evidence that the linkages between Compustat and CRSP have improved over time. But we conclude that there is little reason to believe that this was an issue in the replication. Our analysis is bolstered by the fact that Zuckerman shared two recovered files with us. In the first, all variables were present, but only a small fraction of the observations was recovered. We were able to match approximately 10 percent of our sample. The second recovered data set included four variables, included the Z99 calculation of coverage mismatch, and we were able to match it to 90 percent of our sample. Details of these matching processes, analysis of the match, and full descriptions of the recovered files are in Appendix A2-2.

Careful examination of the first file revealed a likely error in the calculation of the dependent variables in Z99. We discovered that if we (mistakenly) omit debt from the calculation of total value, in interim value in the excess value variable calculation, we are able to much more closely match the Z99 calculated excess values that were recovered and matched to our data set.[13] In diagnostic Procedure 3, reported in Table 4, we use the dependent variable calculated without debt. This produces results similar to Procedure 2. So, this error on its own does not lead to the differences in results.

We did, however, find evidence that the differences between Zacks and I/B/E/S analyst databases likely contribute to our failure to replicate. Using the Z99 coverage mismatch variable from the second recovered file allowed us to repeat the estimation using the original coverage

[13] Specifically, we calculate total capital as equity*1000 + debt, where both equity and debt are measured in *thousands*. That is, this procedure erroneously chops three zeros off the debt and, in practice, has the effect of excluding debt from the total capital calculations. The pair-wise correlations between the excess value measures we calculated in Procedure 2 are 0.71 (sales-based), 0.76 (assets-based), and 0.64 (earnings-based). When omitting debt from the calculation, the correlations for both the assets and sales-based measures increase to 0.95, while for the earnings-based measures it increased to 0.84. See Appendix A2 for further detail and analysis.

mismatch variable based on the Zacks database. In diagnostic Procedure 4, we repeat Procedure 3, but with the Z99 coverage mismatch variable. While these results are weaker than Z99, they are closer to a successful replication. In diagnostic Procedure 5, we use the miscalculated dependent variable together with the Z99 coverage mismatch variable. This combination more closely replicates the Z99 estimates. Our best explanation for these results is that the differences between our results and Z99 are a combination of an error in the calculation of the dependent variables in Z99 and differences between the Zacks and I/B/E/S databases.

We conducted two additional diagnostic tests to understand why differences between Zacks and I/B/E/S may have influenced the results. In Appendix A2-3, we detail that the I/B/E/S database has more systematic analyst coverage for the time period of study, and industries covered by more analysts are more likely, counterintuitively, to produce *higher* coverage mismatch. In Table 2, a comparison of the original Z99 and replication stratifications of industry size based on the 1986 snapshot shows the effect of more comprehensive coverage. Whereas 165 industries had only one to three firms in the Z99 sample that relied on Zacks, in the I/B/E/S based replication, only 116 industries appear this small. In contrast, the Z99 Zacks-based sample has three industries with more than 26 firms, whereas the I/B/E/S sample has 21. Thus, the right shift in the distribution implies, in practice, that more analysts must cover each firm to be considered specialists, especially for firms in industries that moved from the first size category into the next three, or for firms in very large industries. Thus, constructing the Z99 coverage mismatch measure using I/B/E/S leads to fewer analysts classified as specialists, and more covered firms classified as having no specialists, which, in turn, is coded as a mismatch value of 1. Indeed, there are more firm-year observations with a mismatch value of 1 in the replication than in Z99.

To isolate whether the firms coded with coverage mismatch values of 1 in the replicated sample but not in Z99 affect the results, we conducted two additional tests: In diagnostic Procedure 6, we repeat Procedure 2, which uses the correct dependent variable, but we then exclude observations in which coverage mismatch equals 1 (i.e., complete mismatch) in the replicated sample but not the Z99 sample. Diagnostic Procedure 7 is similar to Procedure 6, except we use the incorrectly calculated dependent variable. We summarize these results in Table 3 and report the full descriptive statistics regression results of these estimates in Appendix 1. Using Procedure 6, the replication fails for the sales- and assets-based measures. The coefficients on the earnings-based measure are noisy, and while the replication does not fail the overlap test, these point estimates are all very near zero. When we substitute the incorrectly calculated dependent variable in Procedure 7, we replicate successfully. The assets- and earnings-based models are noisy, but the point estimates are very close to Z99. The estimates pass the overlap test for the sales-based measures, though the sizes of the measured relationship are about 50 percent as large in absolute value. That is, we are able to replicate all three results with the I/B/E/S-based coverage mismatch measure when excluding those firms with coverage mismatch of 1 and using the incorrectly calculated dependent variable. This suggests that the important I/B/E/S-Zacks difference is the more comprehensive coverage of I/B/E/S that leads, counterintuitively, to fewer identified specialists in the Z99 coverage mismatch measure. These differences are sufficiently important, together with the coding error of the dependent variable, to generate the different results between the replication and Z99. To summarize, we conclude that the current replication failed due to a combination of differences in the calculation of excess value due to a coding error in the original, and differences in the underlying analyst databases.

3. EPISTEMIC MAPS

In this section, we use epistemic maps to show estimates under different model assumptions. While Z99 offers 12 specifications to test the theory, we cannot know to what degree the estimates from these specifications are representative of estimates of the larger set of plausible models. The theory in Z99 does not provide specific guidance as to how, precisely, coverage mismatch should be operationalized, nor does it provide strong guidance as to which, if any, control variables should be included in an appropriate test. As such, regardless of whether we can or cannot replicate the specific specifications in Z99, we now ask which different modeling and measurement assumptions map to an economically meaningful negative association between coverage mismatch and excess value, and which do not.

King et al. (2020, p. 45) highlight common forking paths, such as which observations to include in the sample, alternative measures for dependent and independent variables, selection of controls and functional forms, as well as estimation methods and identification strategies. In this present exercise, we focus on robustness to permutations of sampling procedure, measures, and controls (including fixed effects). Each decision can be likened to a fork in the path (Gelman and Loken, 2013). Given an inability to determine which forking path embodies the set of assumptions closest to the truth, Leamer (1985) suggests evaluating *all* paths to understand the distribution of results conditional on these choices, and thereby provide a mapping between different assumptions and results. This mapping approach has been advocated by Simonsohn et al. (2015) in the context of psychology. King et al. (2020) motivate this approach for management research as a partial solution to the philosophy of science problem of testimony. Since authors and readers may not share the same preferences for model choices, and since authors will generally not be able to specify the complete set of judgments that led to particular choices (Wilthold, 2013; and see, for example, the Z99 quote in Footnote 15), systematic mapping between assumptions and results can help the reader to draw their own conclusions about the relationship between variables of interest.

King et al. (2020) recommend publication of results from all models so as to allow the reader to draw their own conclusions based on the set of models they believe most plausible. Following this approach, we publish the results of all regressions on the Open Science Framework (OSF). The specific assumptions we consider are depicted in Figure 1 and described here. We evaluate all plausible permutations of these assumptions.

1. Sample Decisions: We examine Procedures 1, 2, and 2a described above.
2. Dependent Variable: We include the three alternative dependent variables as in Z99.[14]
3. Coverage Mismatch: Industry size is taken into account based on cutoffs specified in Z99 (Table 2). The specialist cutoffs were not dictated by theory but were calibrated to the data.[15] To evaluate the sensitivity of the results to changing these cutoffs, particularly given the more comprehensive coverage in I/B/E/S, we increase and decrease the threshold values in Table 2 by one. Thus, we consider three alternative coverage mismatch measures, which we label the Baseline, Z99 + 1, and Z99 - 1.
4. Controls: We add year fixed effects to the nine controls used in Z99. We treat the year fixed effects as a set (that is, we either include all year fixed effects or none) and, as in Z99, we preclude regressions with both lagged EBIT/Sales and EBIT/Sales. There are

[14] Note that interim values necessary to calculate the dependent variables depend on the choice of sampling procedures 1 or 2. Since Procedure 2a only removes firms from the after the calculation of interim value, the interim values in Procedures 2 and 2a are identical.

[15] Z99 (p. 1417) writes: "[t]o control for industry size, this proportion varies based on the number of covered firms in the industry, as shown in table 1. Note that each of these conditions has been submitted to sensitivity tests, which indicate that the measurement of coverage mismatch is robust across alternative criteria for establishing analyst industry specialization. In addition, informal comparisons of the industry specialties generated by this procedure with those listed in *Nelson's Directory of Investment Research* reveal broad agreement."

383 control variable permutations for each of the 27 DV-IV-Sample combinations. Thus, we run 10,341 regressions for the 1985–1994 time period.

5. Time Period: Evaluating the theory is greatly enhanced by replicating the analysis in a separate time period, which we do by looking at the time period immediately following the Z99 sample: 1995–2004. Thus, we repeat the 10,341 regressions for this later sample.

Figure 1 here

Using a Stata routine developed by one of the authors, we run a firm fixed effects regression with robust standard errors for each of these DV-IV-Sample-Control-Time period permutations.[16] We depict the map between assumptions and results for the 1985–1994 period in Figures 2 and 3. In Figure 2, we stratify by procedure and dependent variable. For example, in the top graph of Figure 2 under the label "Sales," we depict the results of 1,127 regressions estimating the relationship between coverage mismatch and excess value based on sales. For each of the three coverage mismatch variations *baseline, Z99+1,* and *Z99-1* there are 383 possible control variable permutations (3x383 = 1,127). We overlay estimate values and confidence intervals reported in specifications 1–4 above (corresponding to specifications 3–6, 9-12 and 15–18 from Z99), and the bold gray vertical lines are our replication estimates of these models.

Figure 2 here

To facilitate easy interpretation, we depict the implied economic effects of a one SD decrease in coverage mismatch.[17] Thus, the reported quantities are $e^{\hat{\beta}_1 \times -0.32}$-1 where $\hat{\beta}_1$ is the estimated coefficient on coverage mismatch in each regression. The quantities are reported as percentages. For example, in Figure 2, the Z99 result for Model 3 implies a ~5 percent increase in total capital with a single standard deviation decrease in coverage mismatch. The gray areas delimit the 95 percent confidence intervals and the dashed line depicts the point estimates. If a confidence interval includes the bolded Y-axis, then a researcher that had pre-specified the sampling plan, measurement plan, and this specific model would not have not rejected the null hypothesis in a conventional, two-tailed, $p_{crit} = .05,$ frequentist test. The estimated coefficients are ordered by estimate size.

[16] We did not estimate models assuming homoskedasticity of standard errors, though we would expect this to produce smaller standard errors.

[17] 0.32 is the standard deviation of coverage mismatch as calculated by Procedure 2 in the 1985–1994 sample.

Performing the mapping exercise is powerful because it helps researchers determine which assumptions are more likely to be tied to economically meaningful results. Thus, we present the maps by grouping estimates that rely on particular sets of assumptions as depicted in Figure 2. Figure 2 clarifies several results. First, the choice of procedure matters. Under Procedure 1, few combinations of assumptions lead to a measurable negative relationship for the sales-based and assets-based dependent variables. While most combinations of assumptions for the estimates for earnings-based dependent variable are negative, they are relatively small. If we had pre-specified a test, we would not have rejected the null hypothesis that $\hat{\beta}_1 = 0$ for any estimate with excess value base on earnings as the dependent variable. In contrast, in Procedures 2 and 2a, many more of the point estimates imply an increase in firm stock market value as coverage mismatch decreases. In addition, there appears to be little difference between the results of Procedures 2 and 2a.

Second, the choice of dependent variable matters. The estimates for earnings across all three procedures imply the largest effects—a single standard deviation decrease in coverage mismatch implies around a two percent to three percent increase in firm value. However, for the sales- and assets-based measures, many estimates in Procedure 1 imply that a decrease in coverage mismatch would be associated with a small *decrease* in firm value. Under the assumptions of Procedures 2 and 2a, most assumptions imply that a standard deviation decrease in mismatch is associated with less than a one percent change in firm value. For example, if a researcher had pre-specified the models chosen in Z99 (Models 1–4), which in effect is the replication exercise we report in Section 3, they would conclude that there is little evidence of a relationship between coverage mismatch and excess value.

Third, the choice of control variables matters. Different combinations of control variables map to different conclusions about the sign and magnitude of effects. Only if one is willing to assume that only the earnings dependent variable is appropriate might one conclude that the effect is robust to the inclusion of the controls. Moreover, control variable combinations that produce the strongest effects under some dependent variable assumptions do not produce the strongest effects under others. For example, Model (1) produces effects most consistent with the theory when assuming that excess value based on earnings is the appropriate dependent variable, but the least consistent when assuming excess value based on sales or assets is the most appropriate.

Fourth, and relatedly, reporting a subset of estimates may mislead the reader. For example, if only models for earnings using Procedures 2 or 2a were reported, we might conclude that there is evidence of a negative and significant sign—as we would be unaware of the weaker evidence using the other dependent variables, or the results of Procedure 1. The opposite would happen for assets. If we only reported the specifications implied by models predicting assets-based excess value, we would be unaware of a set of assumptions that imply much larger effects.

Finally, we note that drawing conclusions based on the share of estimates in particular regions is not straightforward. Distributions are easily interpretable when draws are equally weighted. However, this is not clearly the case in these graphs. For example, if we equally weight each row in Table 4, then we might say that two-thirds of estimates produce negative coefficients, and one-third produce estimates near 0 with little evidence that they are negative. However, Procedures 2 and 2a are similar relative to Procedure 1, and Procedure 2a may not provide much additional inferential value relative to what we learned from Procedure 2. Thus, if we choose to ignore Procedure 2a, then we would note that only 50 percent of coefficients provide evidence of a negative relationship.

One way to evaluate whether this is the case is to compare the estimates of identical models while varying assumptions. If the results are robust to varying assumptions, we should expect stability in estimates. To do this, we compare models that are identical in every way, except for the sampling procedure. For example, we match the Model 1 estimate (above) from Procedure 1 with the corresponding Model 1 estimates from Procedures 2 and 2a, and similarly, the Model 2 estimates from Procedure 1 with the Model 2 estimates from Procedures 2 and 2a. We see in Table 4 that the procedures lead to very different conclusions: If one believes Model 1 to be the correct specification, then with Procedure 1 the point estimate is 0.00, but with Procedure 2 it is -0.04 and statistically different than zero if there were a properly pre-specified test. Thus, there is inconsistency under the same modeling assumptions.

We can more broadly understand the degree of inconsistency by comparing the results across the three procedures for the 3,447 regression models that were repeated for each procedure. The estimated $\hat{\beta}_1$ coefficients for identical models using Procedures 1 and 2 are highly correlated (0.76), but the precision of these estimates is not (p-values are correlated at -0.17). In over 40 percent of the specifications (1,452/3,447), identical specifications lead to conflicting signs in the estimation of $\hat{\beta}_1$. In contrast, comparing identical models from

Procedures 2 and 2a, the estimated $\hat{\beta}_1 s$ are correlated at 0.99, and corresponding p-values are correlated at 0.88.

We can repeat this exercise, this time conditioning on procedure, and compare models with the same independent variables and controls and different dependent variables. Conditioning on Procedure 1 and the Z99 cutoffs for the coverage mismatch measure, we can compare 191 specifications that are identical on the right-hand side, and differ only in the dependent variables (some models are not comparable across dependent variables because they include lagged dependent variables). The $\hat{\beta}_1$ correlation for identical models using the earnings-based dependent variable and the assets-based dependent variable is -0.39, while the correlation using the assets-based and sales-based measures is 0.65. This is not surprising given that the assets and sales-based measures are highly correlated with each other, and not highly correlated with the earnings-based measure. This indicates that the results are not robust across dependent variables. There are zero cases across these 191 specifications that a researcher would reject the null hypothesis of no effect across all three dependent variables. The sign of $\hat{\beta}_1$ is negative across all three dependent variables in 45 of these 191 specifications. The results are more consistent conditional on Procedure 2. A researcher would reject the null hypothesis in 61 of 191 specifications across all three dependent variables, and the sign is consistently negative in all but one specification. Again, the maps suggest that it is difficult to draw strong conclusions absent a strong prior belief that either Procedure 1 or Procedure 2 is preferable.

Figure 3 reports the same estimates as Figure 2 for Procedures 1 and 2, but we stratify by the calculation of the independent variable-coverage mismatch. We do not depict the results for Procedure 2a because they are little different than Procedure 2. Generally, the map does not appear to be particularly sensitive to changing the cutoffs of the independent variables. Moreover, the results are quite consistent for similar models that vary only in the assumption of independent variable calculation. For example, models that use the Z99 cutoff criterion are correlated with identical models that use the "Z99-1" criterion at 0.98 and with identical models that use the "Z99+1" criterion at 0.91. This consistency does not indicate support for the theory, however. The sign of the estimated $\hat{\beta}_1$ changes in 20 percent of the 3,447 triplets of identical models. In this case, the map suggests that one particular set of assumptions, the cutoff points for the measurement of coverage mismatch, is not that important.

Figure 3 here

We now turn to the 1995–2004 sample. To the extent that one believes that the later period represents a new draw from the same population, this could be interpreted as an exact replication. To maintain this belief, one must believe that the later time period has not changed the data generating process. We replicate Figures 2 and 3 for the 1995–2004 sample in Figures 4 and 5. The coefficients for the Z99 reported models and their estimated Procedure 2 equivalents are included for reference. Comparing the Z99 coefficients to the results in Figure 4, we see that the Z99 estimates suggest much larger effects than those we estimated in this later period. This result is very similar to the comparison of Z99 estimates to our estimates using the 1985–1994 period. Interestingly, for the excess value based on earnings measures, the confidence intervals of our estimates and Z99's overlap, suggesting that we cannot reject the null hypothesis that our estimates and Z99s are different from each other. However, this alone does not indicate support for the theory, as the dashed line reveals that many of the point estimates we estimated for Models 1–4 associate a negative change in earnings-based excess value with a decrease in coverage mismatch.

Figures 4,5 here

More generally, we see again that the conclusions one draws from the map will be highly sensitive to beliefs about the correct choice of model and sampling procedure. For example, many of the estimates using Procedure 2 and the earnings-based measure suggest a negative association between decreased coverage mismatch and excess value, while Procedure 1 using the sales-based measure suggests a more positive association. The pairwise correlations between $\hat{\beta}_1$ estimates across the Procedures 1 and 2, 1 and 2a, and 2 and 2a respectively are 0.20, 0.64, and 0.75. In 58 of the 191 model triplets, the sign of the point estimate is not consistent.

One should interpret this instability as evidence of sample variation. We see similar instability when comparing estimates from the 1985–1994 period to those from 1995–2004. Across the 10,341 assumption pairs (i.e., paired procedures, independent and dependent variables and controls), $\hat{\beta}_1$ estimates are *negatively* correlated at 0.34, and 36 percent of pairs yield opposing $\hat{\beta}_1$ signs. In only 599 of the 10,341 specifications would one reject the properly pre-specified null hypothesis that there is no effect, creating identical assumptions across the two time periods. To illustrate the negative correlation result, in Figure 6, we take the models that produce the 200 most negative $\hat{\beta}_1$ estimates from the early period and map them in the later period. Note that in this figure we depict the coefficient estimates as opposed to the effect magnitudes, thus the most negative estimates are those most consistent with the theory. We find that the assumptions that produce the most negative estimates, and are thus consistent with the theory in the 1985–1994 period, produce many estimates that are positive in the later period. Thus, to interpret the findings as evidence in support of the theory, one must believe that the pairing exercise is misguided due to the unsuitability of either the earlier or later period for the purposes of this empirical exercise. This variance in results across time periods does not lead to the conclusion that the theory is not true, but simply that there is not strong evidence in its favor. But we can say that if the theory is true, in these data, the mechanism is too weak relative to other factors to generate consistent patterns across alternative sampling schemes, measurement decisions, and time periods. In any case, we cannot draw a conclusion directly with these data that there is robust evidence consistent with the theory.

Figure 6 here

4. DISCUSSION

Our analysis leads to three central conclusions. First, we are unable to replicate Z99s results in terms of economic significance. We can reproduce the signs of the estimated coefficients using Procedure 2, though when we do, the economic significance is diminished. Second, the results are sensitive to the modeling and sample construction procedures, and not generally robust across dependent variables. Epistemic mapping shows that almost any result can be obtained based across various combinations of control variables, predictor and outcome variable construction, and samples. Third, these general conclusions are true within both the 1985–1994 and 1995–2004 samples. Moreover, the assumptions that lead to the more negative coefficient estimates in the 1985–1994 sample are not the same assumptions that lead to the more negative coefficients in the 1995–2004 sample, and vice versa.

We believe that the Z99 empirical results provide insufficient evidence to make managerial recommendations about category management. The results are highly sensitive to the choice of dependent variables, controls, sample construction procedure, and period of study. In this case, the theory does not provide strong guidance as to which choice to make and interpreting the statistics as a test of the theory requires making such a choice. If we had found consistent results in support of the theory across these assumption choices, we might have concluded that there is stronger evidence of the predictive value of the theory. But the inconsistency of results leads us to believe that there is no evidence in these data, in either period, of a systematic pattern between coverage mismatch and capital market value.

While there is little evidence coverage mismatch is related to excess value in our data, this analysis only examines one potential approach to measurement, and some parts of the map are left out. For example, recent work by Hoberg and Phillips (2016) indicates that SIC industry codes are imperfect measures of groups of relevant competitors. If this is true, the coverage mismatch measure may not be the best way to measure deviations from expected categories because SIC codes may not accurately reflect the cognitive categories of investors. Indeed, this critique is implicit in Zuckerman's own follow-on work (2004), in which he measures "structural coherence" based on the degree to which different analysts tend to cover the same firms. The measure relies on the degree to which analysts themselves follow the same firms, and Zuckerman therefore does not rely on the third-party SIC classifications. However, lending credibility to the original measure, Zuckerman (2000) also documents a positive relationship between coverage mismatch and segment exit in the same sample as Z99, and Feldman (2016) reports a positive association with coverage mismatch and analyst forecast errors.

Our exercise highlights two points. First, it was difficult for us, as skeptical readers, to evaluate the generality of conclusions from Z99. The analysis was not pre-specified, Indeed, in most archival studies, pre-specification is often very difficult. Hence, we did not have the information needed to interpret the Z99 reported estimates of precision (i.e., standard errors or p-values). Second, based on our experience with this project, replication exercises which might clarify the reliability of results are costly and difficult. Our replication was challenged by changing data sources and or data availability, and incomplete or ambiguous testimony of methods. Second, in order to evaluate which assumptions in Z99 mattered for the results, we needed to create a systematic, epistemic map.

We offer three complementary remedies. First, if data and code were published, this would allow readers to reproduce results, and better understand the robustness of reported statistics to alternative specifications. Such transparency may improve readers' confidence in the originally reported statistics. When publishing data is not possible, a similar effect could be achieved if authors retained copies of their original data. This would better allow future scholars to understand if database changes made by commercial vendors affect our inference. Second, in this study, epistemic maps were a powerful diagnostic tool that helped our understanding of the robustness of results across an expansive number of alternative assumptions (King et al., 2020). The publication of such maps as a matter of course may stem debates about the reliability and sensitivity of results. This transparency would complement other remedies, such as depicting relationships across the support of independent variables using graphs (e.g., binned scatterplots (Starr and Goldfarb, 2020)). In this study, mapping allowed us to evaluate empirical support for the underlying theory more critically, as it focused our attention on the robustness across different assumptions and samples. We hope that this will be true in other studies as well. Third, without pre-specification, researcher credibility may be enhanced with weaker claims to generalizability that empower the reader to more actively engage and draw their own conclusions. Given that it is difficult to know the degree in which one's results are driven by sample variation, modest claims may actually facilitate greater trust with the reader (King et al., 2020). We believe that more expansive and transparent reporting, as well as more modest claims, will be helpful to developing a more robust cumulative body of knowledge in strategy research.

5. CONCLUSION

We replicate "*The Categorical Imperative: Securities Analysts and the Illegitimacy Discount"* Zuckerman (1999) and conduct an extensive robustness analysis of the empirical exercise. Our results are inconsistent with those reported in Z99 for two of the three sets of dependent variables. We find that part of the difference is likely due to coverage differences between the Zacks and I/B/E/S databases, and part of the difference is, most likely, due to an error in the original calculations. Some of our estimates are consistent with the theory, but others are not. Those results that are consistent suggest effect sizes that are much smaller than those reported in Z99.

To better understand the sensitivity of the results, we build a series of epistemic maps in which we systematically map various sample construction, sample inclusion, and measurement assumptions to the estimated relationship between coverage mismatch and excess value. We find that the results are highly sensitive to these assumptions and infer that these differences are likely due to sample variation. We repeat the exercise using a later sample (1995–2004) and draw similar conclusions. Results are sensitive to sample construction, measurement, and model choice, not only within the 1995–2004 sample. Assumptions associated with estimates consistent with the theory in the early period do not map to estimates consistent with the theory in the later period, and vice versa. In this sense, our results bring into question the internal and the external validity of the hypothesis tested in Zuckerman (1999).

Our results do not, and indeed *cannot,* suggest Zuckerman's theory, or any theory of categorization is not useful or true. We can only conclude that the results of this specific empirical exercise, which relates specialist coverage measured to excess value, are fragile and that they provide insufficient evidence to make managerial recommendations. We hope that this approach, of systematically evaluating the mapping of assumptions to results will be more widely adopted in Strategic Management Research.

References

–Benner, M.J. and Ranganathan, R., 2012. Offsetting illegitimacy? How pressures from securities analysts influence incumbents in the face of new technologies. *Academy of Management Journal*, *55*(1):213–233.

Bettis, R.A., Ethiraj, S., Gambardella, A., Helfat, C. and Mitchell, W., 2016. Creating repeatable cumulative knowledge in strategic management. *Strategic Management Journal*, *37*(2):257–261.

Bettis RA, Helfat CE, Shaver JM. 2016. The necessity, logic, and forms of replication. *Strategic Management Journal* **37**(11): 2193–2203.

Berger, P.G. and Ofek, E., 1995. Diversification's effect on firm value. *Journal of financial economics*, *37*(1):39–65.

Durand, R., Rao, H. and Monin, P., 2007. Code and conduct in French cuisine: Impact of code changes on external evaluations. *Strategic Management Journal*, *28*(5):455–472.

Feldman, E.R., 2016. Corporate spinoffs and analysts' coverage decisions: The implications for diversified firms. *Strategic Management Journal*.

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'fishing Expedition' or 'p-Hacking' and the Research Hypothesis Was Posited ahead of Time." Department of Statistics, Columbia University. https://osf.io/n3axs/download.

—Granqvist, N., Grodal, S. and Woolley, J.L., 2013. Hedging your bets: Explaining executives' market labeling strategies in nanotechnology. *Organization Science*, *24*(2):395–413.

Hoberg G., and Phillips, G., 2016. Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy* 124 (5):1423–1465.

Hsu, G., 2006. Jacks of all trades and masters of none: Audiences' reactions to spanning genres in feature film production. *Administrative Science Quarterly*, *51*(3):420–450.

Khaire, M. and Wadhwani, R.D., 2010. Changing landscapes: The construction of meaning and value in a new market category—Modern Indian art. *Academy of Management Journal*, *53*(6):1281–1304.

King, A.A., Goldfarb, B., and Simcoe T. S. 2020. Learning from Testimony on Quantitative Research in Management", *Academy of Management Review (forthcoming).*

Leamer, E. E. 1983. "Chapter 5 Model Choice and Specification Analysis." In *Handbook of Econometrics*, 1:285–330. Elsevier.

Litov, L.P., Moreton, P. and Zenger, T.R., 2012. Corporate strategy, analyst coverage, and the uniqueness paradox. *Management Science*, *58*(10):1797–1815.

Lo, J.Y.C. and Kennedy, M.T., 2014. Approval in nanotechnology patents: Micro and macro factors that affect reactions to category blending. *Organization Science*, *26*(1):119–139.

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Navis, C. and Glynn, M.A., 2010. How new market categories emerge: Temporal dynamics of legitimacy, identity, and entrepreneurship in satellite radio, 1990–2005. *Administrative Science Quarterly*, *55*(3):439–471.

Navis, C. and Glynn, M.A., 2011. Legitimate distinctiveness and the entrepreneurial identity: Influence on investor judgments of new venture plausibility. *Academy of Management Review*, *36*(3):479–499.

Negro, G. and Leung, M.D., 2013. "Actual" and perceptual effects of category spanning. *Organization Science*, *24*(3):684–696.

–Paolella, L. and Durand, R., 2016. Category spanning, evaluation, and performance: Revised theory and test on the corporate law market. *Academy of Management Journal*, *59*(1):330–351.

Pontikes, E.G., 2012. Two sides of the same coin: How ambiguous classification affects multiple audiences' evaluations. *Administrative Science Quarterly*, *57*(1):81–118

Wry, T., Lounsbury, M. and Jennings, P.D., 2014. Hybrid vigor: Securing venture capital by spanning categories in nanotechnology. *Academy of Management Journal*, *57*(5):1309–1333.

Sala-i-Martin. X., 1997. "I Just Ran Two Million Regressions." *The American Economic Review* 87 (2): 178–83.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*(3), 182–186.

Simonsohn, U., Simmons, J.P., and Nelson, L. D. 2015. "Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications." https://doi.org/10.2139/ssrn.2694998.

Smith, E.B., 2011. Identities as lenses: How organizational identity affects audiences' evaluation of organizational performance. *Administrative Science Quarterly*, *56*(1):61–94.

Spanos, Aris. 2014. "Recurring Controversies about P Values and Confidence Intervals Revisited." *Ecology*, 95(3): 645-651

Starr, Evan, and Brent Goldfarb. 2020. "Binned Scatterplots: A Simple Tool to Make Research Easier and Better." *Strategic Management Journal*, June. https://doi.org/10.1002/smj.3199.

Vergne, J.P. and Wry, T., 2014. Categorizing categorization research: Review, integration, and future directions. *Journal of Management Studies*, *51*(1):56–94.

Wilholt, T. (2013). Epistemic trust in science. The British Journal for the Philosophy of Science 64(2): 233–253

Zuckerman, E.W., 1999. The categorical imperative: Securities analysts and the illegitimacy discount. *American journal of sociology*, *104*(5):1398–1438.

Zuckerman, E.W., 2000. Focusing the corporate product: Securities analysts and de-diversification. *Administrative Science Quarterly*, *45*(3):591–619.

Zuckerman, E.W., 2004. Structural incoherence and stock market activity. American Sociological Review, 69(3): pp.405–432.

TABLES

TABLE 1 Replication Procedures

	Procedures							
	Replications			Diagnostics				
	1	**2**	**2a**	**3**	**4**	**5**	**6**	**7**
Exclude firms with any segment in finance industry, non-US firms, and firms that are not traded on major exchanges from the initial sample	Y							
Include firms with any segment in finance industry, non-US firms, and firms that are not traded on major exchanges from the initial sample, and only exclude them in the final analysis		Y	Y	Y	Y	Y	Y	Y
Include firms that have sales less than $20 million	Y							
Exclude firms that have sales less than $20 million in the construction of DVs (these firms will have missing values in DVs)		Y	Y	Y	Y	Y	Y	Y
Take square root of Herfindahl index		Y	Y	Y	Y	Y	Y	Y
Total capital (in millions) = equity + debt	Y	Y			Y		Y	
Total capital (in thousands) = equity * 1000 + debt			Y	Y		Y		Y
DVs from the recovered file in the final analysis			Y		Y	Y		

Description	Per Z99 description	Best guess of Z99 implementation	2nd best guess of Z99 implementation	Best guess of Z99 implementation with the incorrect DVs	Best guess with the correct DVs and the coverage mismatch from the recovered file	Best guess with the incorrect DVs and the coverage mismatch from the recovered file	Procedure 2, using the sample overlapped with the recovered file and excluding observations that have mismatch value of '1' only in the replication but not in the original sample	Procedure 3, using the sample overlapped with the recovered file and excluding observations that have mismatch value of '1' only in the replication but not in the original sample

Note: Z99 specifies only using firms with sales more than $20 million in the calculation of median ratios by industry. This criterion doesn't automatically exclude firms that have sales less than $20 million from the analysis sample. Z99 does not indicate that these firms will have missing values in the dependent variables. We are unsure whether these firms are excluded from the whole sample, or just the sample for the construction of the dependent variables.

Table 2 (Table 1 in Zuckerman, 1999; Original + Replication)

Z99 Table 1			
No. of Covered Firms with Highest Proportion of Sales in the Industry	Minimal Proportion for Analyst Coverage	No. of Relevant Three-Digit Industries, 1986 (Z99)	No. of Relevant Three-Digit Industries, 1986 (Replicated)
1–3	1	165	116
4–5	0.8	30	31
6–10	0.6	36	54
11–15	0.5	9	27
16–20	0.4	9	11
21–25	0.3	5	6
26+	0.2	3	21

Table 3 Summary Statistics

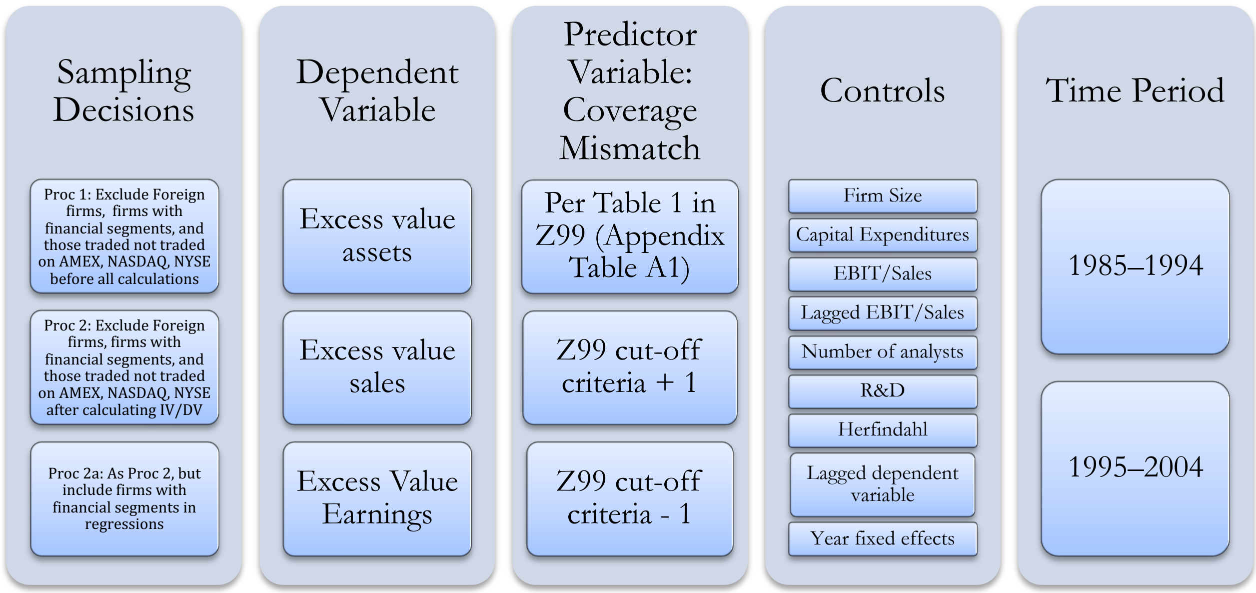
Variables	Obs (firm year)		Mean			S.D.			Min			Max		
	Original	Replicated	Original	Replicated		Original	Replicated		Original	Replicated		Original	Replicated	
				P3	P2		P3	P2		P3	P2		P3	P2
Excess value-sales based	27,597	27,945	-0.13	-0.08	-0.03	1.04	0.90	0.67	-16.31	-6.60	-4.97	4.2	4.06	3.64
Excess value-assets based	26,346	26,569	-0.12	-0.08	0.01	0.96	0.82	0.52	-16.16	-5.50	-4.53	3.71	3.67	3.05
Excess value-EBIT based	24,652	24,659	0.12	-0.05	0.01	1.02	1.23	1.09	-16.27	-9.57	-7.45	9.7	9.70	9.58
Log of assets	31,602	28,933	5.12	5.07		1.69	1.67		0.06	-1.53		11.43	11.43	
EBIT/sales	31,602	28,848	0.07	0.07		0.14	0.13		-3.92	-3.49		0.78	0.78	
Capital expenditure/sales	31,116	28,492	0.09	0.08		0.19	0.19		0	0		8.31	8.31	
Herfindahl	31,600	28,933	0.92	0.92		0.14	0.14		0.34	0.00		1	1	
R&D/sales	15,945	14,986	0.05	0.05		0.09	0.09		0	0		3.41	3.41	
Number of analysts	31,602	28,933	6.26	6.93		8.33	9.24		0	0		62	66	
Coverage mismatch	31,526	28,933	0.73	0.82		0.35	0.32		0	0		1	1	

Table 4 Coverage Mismatch Estimates Across Procedures 1985–1994

	DV: Excess value - sales based				DV: Excess value - assets based				DV: Excess value - EBIT based			
Pro-cedure	M3	M4	M5	M6	M9	M10	M11	M12	M15	M16	M17	M18
						Panel A: Replication						
Z99	-0.15	-0.14	-0.13	-0.09	-0.11	-0.12	-0.11	-0.09	-0.11	-0.07	-0.05	-0.08
	[-0.22,-0.09]	[-0.19,-0.09]	[-0.18,-0.08]	[-0.14,-0.05]	[-0.17,-0.05]	[-0.17,-0.08]	[-0.16,-0.07]	[-0.14,-0.05]	[-0.20,-0.01]	[-0.13,-0.00]	[-0.12,0.01]	[-0.15,-0.01]
1	0.00	0.01	0.01	0.02	0.04	0.00	0.01	0.00	-0.07	0.00	0.01	0.00
	[-0.07,0.07]	[-0.03,0.06]	[-0.03,0.05]	[-0.01,0.06]	[-0.01,0.09]	[-0.03,0.04]	[-0.02,0.04]	[-0.03,0.03]	[-0.21,0.07]	[-0.08,0.09]	[-0.08,0.09]	[-0.09,0.09]
2	-0.04	-0.02	-0.01	0.00	-0.03	-0.04	-0.03	-0.03	-0.10	-0.07	-0.06	-0.08
	[-0.07,-0.01]	[-0.04,0.01]	[-0.04,0.01]	[-0.03,0.02]	[-0.06,-0.03]	[-0.06,-0.02]	[-0.05,-0.01]	[-0.05,-0.01]	[-0.19,-0.01]	[-0.14,-0.01]	[-0.12,0.01]	[-0.15,-0.01]
2a	-0.04	-0.02	-0.01	0.00	-0.03	-0.03	-0.03	-0.02	-0.09	-0.06	-0.04	-0.06
	[-0.07,-0.01]	[-0.04,0.01]	[-0.04,0.01]	[-0.03,0.02]	[-0.06,0.00]	[-0.06,-0.02]	[-0.05,-0.01]	[-0.04,-0.00]	[-0.18,-0.00]	[-0.12,0.00]	[-0.10,0.02]	[-0.12,0.01]
						Panel B: Diagnostics						
3	-0.03	-0.01	-0.01	0.01	-0.03	-0.03	-0.03	-0.01	-0.10	-0.07	-0.05	-0.07
	[-0.07,0.02]	[-0.05,0.02]	[-0.04,0.03]	[-0.02,0.05]	[-0.08,0.01]	[-0.06,-0.00]	[-0.06,0.01]	[-0.04,0.02]	[-0.20,-0.00]	[-0.14,-0.00]	[-0.12,0.02]	[-0.15,-0.00]
4	-0.11	-0.06	-0.07	-0.03	-0.10	-0.09	-0.09	-0.05	-0.06	0.02	0.03	0.03
	[-0.15,-0.07]	[-0.09,-0.04]	[-0.10,-0.04]	[-0.06,-0.00]	[-0.14,-0.06]	[-0.11,-0.06]	[-0.12,-0.07]	[-0.07,-0.03]	[-0.18,0.06]	[-0.06,0.1]	[-0.05,0.12]	[-0.05,0.12]
5	-0.18	-0.12	-0.12	-0.05	-0.18	-0.14	-0.14	-0.07	-0.15	-0.03	-0.02	-0.02
	[-0.24,-0.13]	[-0.16,-0.08]	[-0.16,-0.08]	[-0.08,-0.01]	[-0.23,-0.12]	[-0.18,-0.10]	[-0.18,-0.10]	[-0.10,-0.04]	[-0.28,-0.02]	[-0.12,0.05]	[-0.10,0.07]	[-0.11,0.07]
6	-0.02	0.00	0.00	0.02	-0.04	-0.06	-0.05	-0.03	0.00	0.01	0.02	-0.01
	[-0.08,0.04]	[-0.04,0.04]	[-0.04,0.05]	[-0.02,0.06]	[-0.09,0.01]	[-0.09,-0.02]	[-0.09,-0.02]	[-0.07,-0.00]	[-0.18,0.17]	[-0.10,0.12]	[-0.09,0.13]	[-0.13,0.11]
7	-0.09	-0.07	-0.06	-0.01	-0.12	-0.12	-0.11	-0.06	-0.11	-0.08	-0.05	-0.09
	[-0.17,-0.01]	[-0.12,-0.01]	[-0.12,-0.00]	[-0.06,0.04]	[-0.12,-0.05]	[-0.17,-0.07]	[-0.16,-0.05]	[-0.11,-0.01]	[-0.29,0.07]	[-0.20,0.04]	[-0.17,0.07]	[-0.22,0.04]
Z99	-0.15	-0.14	-0.13	-0.09	-0.11	-0.12	-0.11	-0.09	-0.11	-0.07	-0.05	-0.08
	[-0.22,-0.09]	[-0.19,-0.09]	[-0.18,-0.08]	[-0.14,-0.05]	[-0.17,-0.05]	[-0.17,-0.08]	[-0.16,-0.07]	[-0.14,-0.05]	[-0.20,-0.01]	[-0.13,-0.00]	[-0.12,0.01]	[-0.15,-0.01]

FIGURES

FIGURE 1 Critical forking paths



Note: Details of Procedure 1, 2, and 2a are specified in Table 1.

FIGURE 2 Epistemic Maps: 1985-1994

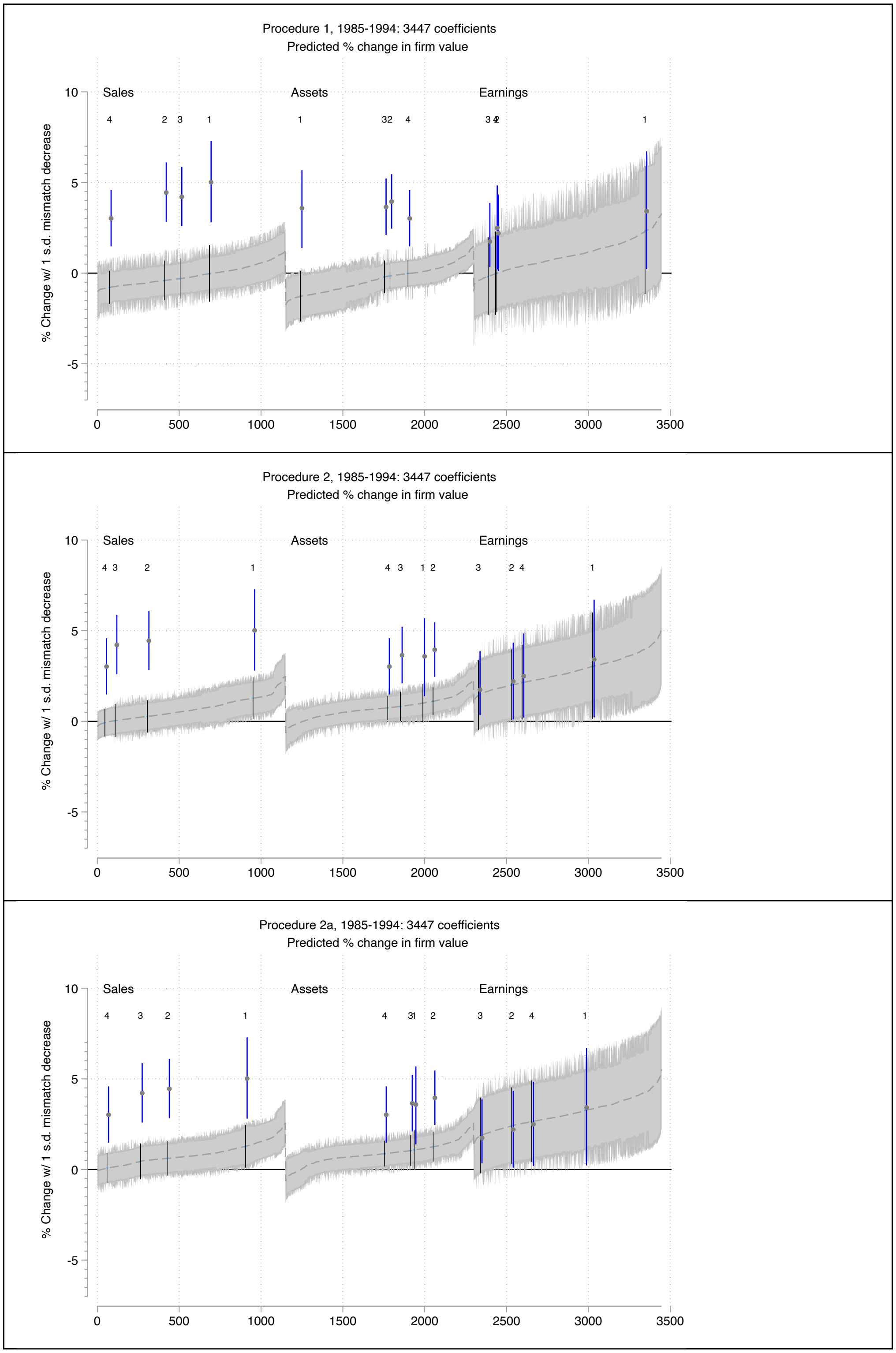


FIGURE 3 Epistemic Maps 1985-1994, Varying Coverage Mismatch Criteria

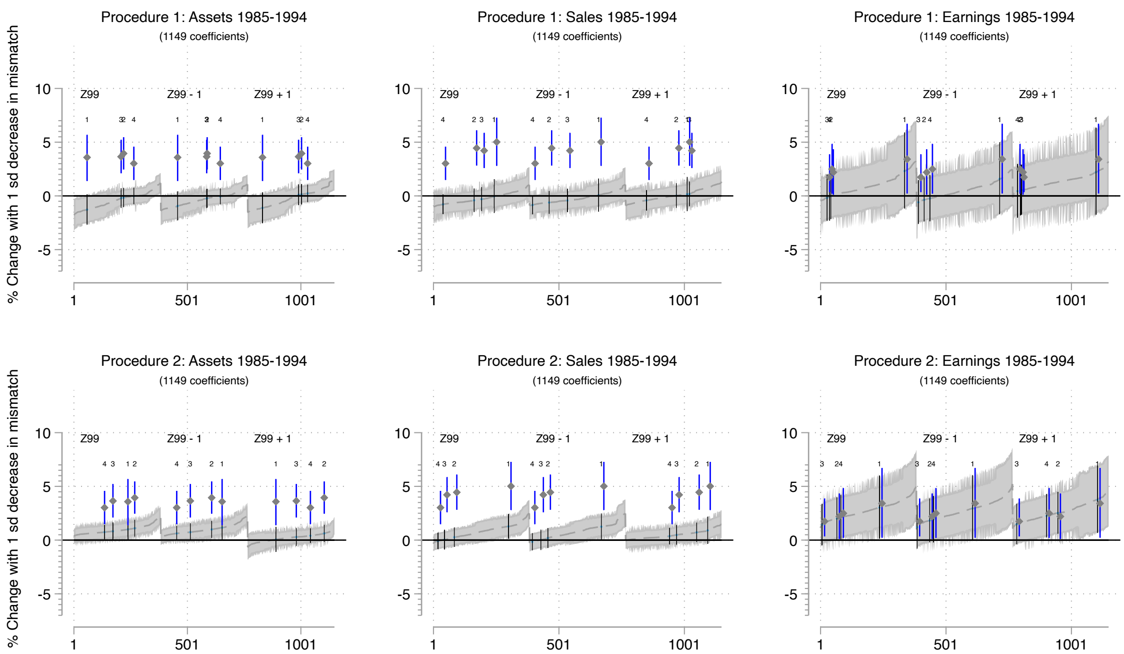
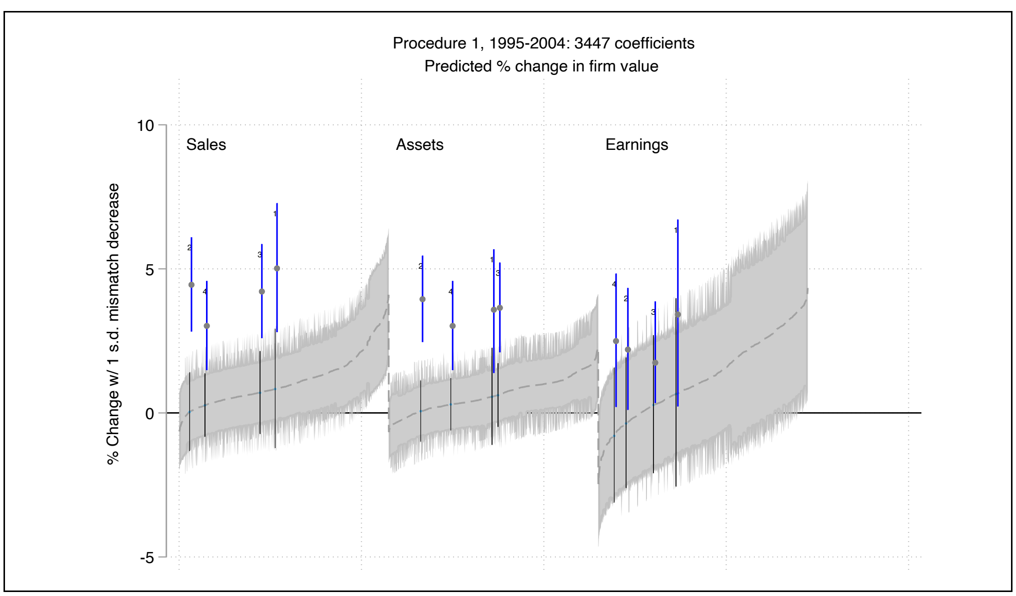


FIGURE 4 Epistemic Maps, 1995-2004



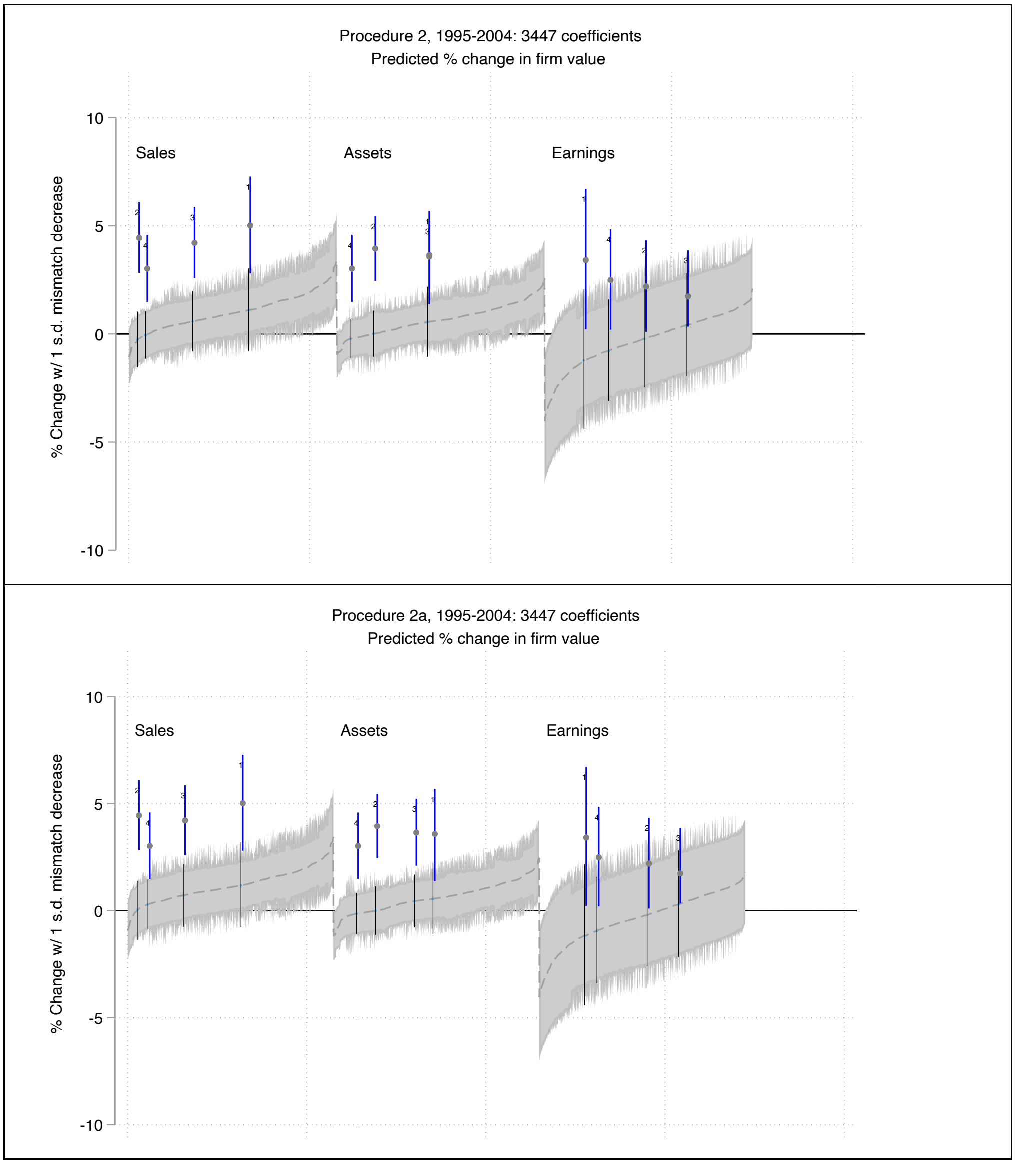


FIGURE 5 Epistemic Maps 1995-2004 Varying Coverage Mismatch Criteria

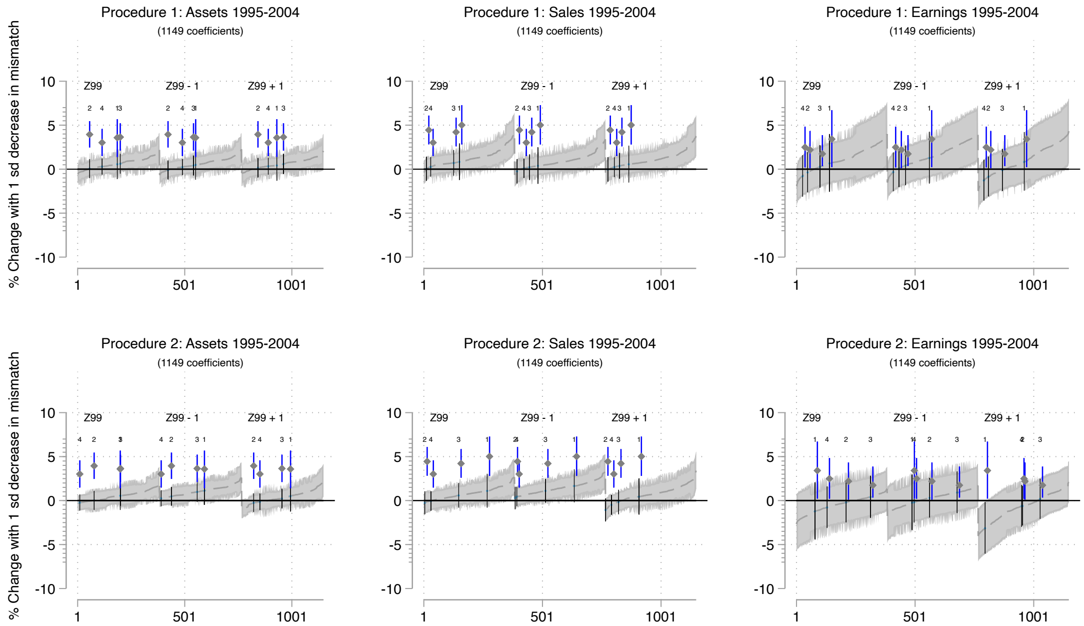
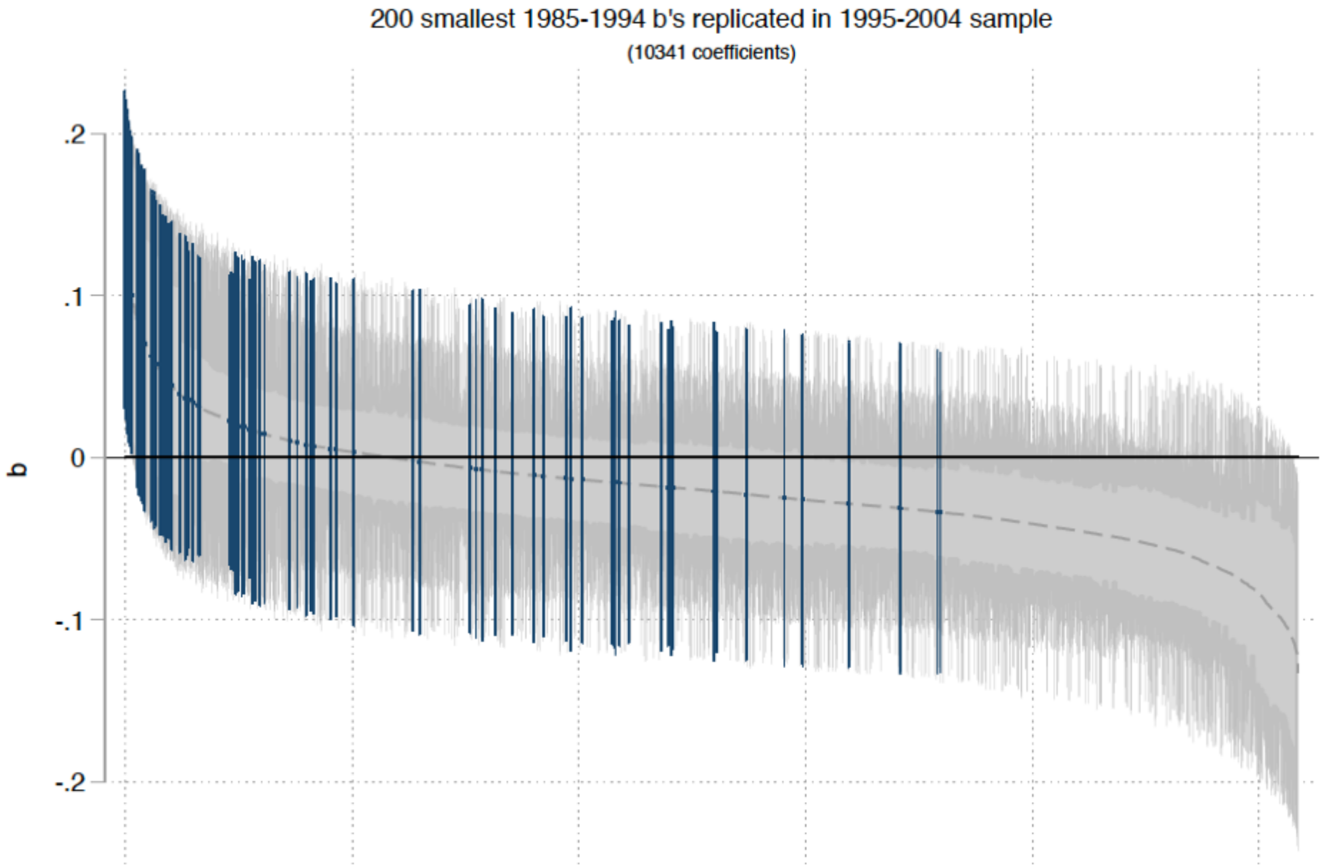


FIGURE 6 Instability of model results across early and later samples


200 smallest 1985-1994 b's replicated in 1995-2004 sample
(10341 coefficients)
b
.2
.1
0
-.1
-.2