The Stock Market, Product Uniqueness, and Comovement of Peer Firms

Gerard Hoberg and Gordon Phillips*

November 14, 2012

ABSTRACT

Using text-based computational analysis, we examine whether firm stock market valuations reflect product uniqueness and how firms covary with peers in the stock market. We find that firms have higher stock market valuations than matched peer firms when their product portfolios are more unique relative to matched peer firms. We also find that the returns of text-based peers better explain own-stock returns than traditional industry peers. These results hold for both conglomerate firms and single segment firms using best-matched single-segment peer firms. Overall, our findings show that the stock market values product uniqueness and recognizes peer groups based on fundamental product characteristics that are not reflected in standard peer groupings.

*University of Maryland and University of Southern California and National Bureau of Economic Research, respectively. Hoberg can be reached at ghoberg@rhsmith.umd.edu and Phillips can be reached at gordon.phillips@marshall.usc.edu. We thank seminar participants at Duisenberg School of Finance and Tinbergen Institute, Erasmus University, Rotterdam School of Management, Stanford University, Tilberg University, University of Chicago, University of Illinois and the University of Mannheim for helpful comments. All errors are the authors alone.

A fundamental question both in corporate finance and asset pricing is whether and how the stock market uses information from peer firms. The analysis of the effects of peer firms has focused on the effect of peers on capital structure (MacKay and Phillips (2005), Leary and Roberts (2010) and Rauh and Sufi (2010)) and on industry momentum in asset pricing by Moskowitz and Grinblatt (1999). It remains unknown whether and how the stock market values and incorporates peer firm fundamentals in affecting the valuation and stock returns of firms.[1]

Using new text-based identification of peer firms that best replicate the product offerings of the firm, we show that the stock market does indeed account for information associated with peer firms. We examine how the stock market uses information from peer firms in two ways. First, using peer firms that are matched using text based replication methods, we examine whether firms that have more unique products relative to their peers have higher stock market valuations. Second, we examine whether text-based peer firms explain stock market comovement more than standard industry peer groupings, and whether firms with peers whose products more closely match theirs have higher stock-market comovement.

We find that text-based replication peers significantly outperform traditional industry peers in explaining firm valuations and peer comovement in the stock market. Firms have higher stock market valuations than replication peer firms when their product offerings are more unique relative to the replicating peers. These higher valuations also are long-lasting, as firms with more unique products do not experience ex-post abnormal return reversals. Firm returns also covary more with text-based peers than they do with peers using standard industry classifications. These results hold for both conglomerate firms and single segment firms. Overall, our findings show that the stock market values product uniqueness and recognizes peer groups based on fundamental product characteristics that are not reflected in standard industry groupings.

Our paper extends previous identification of competitors using text-based analysis by Hoberg and Phillips (2010a) by assigning different weights to peer firms based on

[1]Additional studies on the effects of peer firm financial policies on real decisions in a strategic context include Phillips (1995), Chevalier (1995), and Khanna and Tice (2000).

how each uniquely contributes to creating a near replica (on the basis of both product offerings and accounting characteristics) of a given firm under consideration. For example, potential peers of Apple would include Dell, HP, Samsung, Motorola, Microsoft and Sony (which owns Columbia music). Our procedure identifies a weighted portfolio of replication peers that best reconstructs Apple's actual product offerings and accounting characteristics. Peers getting the highest weights (A) are most similar to Apple in their product offerings and (B) are unique in the aspects of Apple's product description that they replicate. This reconstructive approach is particularly insightful for understanding conglomerate industry structure (and conglomerate valuation) given the more complex product offerings of conglomerates. Although we generate replicas of all firms, we only include single-segment firms as reconstructive peers to be able to benchmark valuations as in previous studies by Lang and Stulz (1994) and Berger and Ofek (1995).[2]

Although numerous papers examine links between competition and firm decisions, little has been done to examine whether (or the extent to which) the stock market values a firm's differentiation from potential rivals. For all firms - both single-segment and multiple-segment conglomerate firms - we find that firm valuations are higher when the firm has less perfectly matched replicating peers using firm product descriptions. In contrast, firms having product descriptions and characteristics that are better matched or "spanned" by peer firms, trade at discounts. The results are consistent with firms that are difficult to replicate having unique products and facing less direct competition now, and also for the foreseeable future (a poor fit using peer firms indicates both that direct rivals do not exist now and also that nearby firms cannot at a later date exchange assets to replicate the given firm). These results are consistent with the stock market valuing product uniqueness.

Our valuation results for our product uniqueness measure captured by the difficulty to replicate are consistent with the conclusion that it is not easy for a competitor to introduce similar successful products. The case of Apple Inc. versus its peers is illustrative. Apple's peers, Dell and HP have both tried to introduce successful tablet

[2]These "network" benchmarks represent best matches in the product market analogous to a weighted Facebook circle of friends (both close friends and acquaintances).

computers, while Sony and Microsoft have introduced new digital music players. Apple still has very high margins and market shares in each of these markets multiple years after first introducing its products - despite efforts by peers to replicate Apple's successful product offerings.

We use these new benchmarks and measures of product uniqueness to examine cross-sectional differences in conglomerate and single-segment valuations. For conglomerate firms, we also show that our weighted benchmarks provide economically large improvements relative to existing methods in their ability to accurately match conglomerate valuations and characteristics. Although we do not focus on the conglomerate discount, which many studies document to be a result of self-selection, our framework shows that it also disappears using our replicating peers.[3]

We also show that our text based replicating peers significantly covary more in the stock market than do traditional industry peers. This comovement is particularly high when firms have more closely matched replicating peers, which have more similar products. These firms, that are easier to replicate using text peers, thus comove more with their peers than firms that are more difficult to replicate. These findings are consistent with firms with similar products moving more tightly together given that firms with similar products face similar cost and demand shocks. Our new text-based peers are significantly more important both in magnitude and significance than traditional SIC-based peers in explaining industry comovement.

We further consider both simultaneous stock return comovement (a test of importance), whether lagged peer returns predict future own-firm returns (a test of how information is conveyed in the stock market), and whether results differ for systematic and idiosyncratic risk. Simultaneous return tests indicate that the stock market strongly reflects common information in the competitive environment, and that these effects are better measured using replicating text-based peers. The lagged

[3]For articles on the average or median discount of conglomerate firms see Wernerfelt and Montgomery (1988), Lang and Stulz (1994), Berger and Ofek (1995), Comment and Jarrell (1995), Servaes (1996), Lins and Servaes (1999), and Lamont and Polk (2002) find evidence of a diversification discount. This average discount has been shown to be related to self-selection by Campa and Kedia (2002), Graham, Lemmon, and Wolf (2002), and Villalonga (2004b) and by data problems by Villalonga (2004a) and merger accounting by Custodio (2010). See Maksimovic and Phillips (2007) for a detailed survey.

return tests indicate that additional information from replicating peers continues to disseminate into own-firm returns over time, suggesting that the market initially under-reacts to information from replicating peers. These results are economically large, and in some cases, these measures are as much as five times more influential as compared to measures using SIC-based peers. We also find that simultaneous co-movement is strongest for systematic return shocks, but that lagged return spillovers are strongest for idiosyncratic return shocks. We conclude that the market is more informationally efficient regarding systematic shocks than idiosyncratic shocks.

Our paper makes three main contributions. Our first contribution is to examine the link between product market variables and stock market valuations in cross section. For both single-segment and conglomerate firms, we find that firm valuations are higher when the firm is more difficult to replicate using matched single-segment firms, and when its segments have high-value industries between them. Given that our focus is on understanding the valuation of firms in cross section - both single-segment and conglomerate firms, it fills a gap empirically that Stein (2003) identifies in his survey paper.[4]. In all, our findings support theoretical links between and valuation and product uniqueness.

Our second contribution is to show that our text-based replicating peers are significantly more important both in magnitude and significance than traditional industry groupings in explaining peer comovement. We document the extent to which industry comovement and return predictability are based on product fundamentals.

Our third contribution is methodological: we present new text-based methods that use constrained optimization to generate single-segment benchmarks for both single-segment firms and conglomerate firms. We also document how these replicating peers can be identified using a closed form solution to a constrained optimization problem. These benchmarks offer significant gains in accuracy relative to existing methods and we show that these groupings covary together in the stock-market.

[4]Stein (2003) writes the focus should be on "under what conditions is an internal capital market most (or least) likely to add value relative to an external capital markets benchmark?" Our paper addresses this question conceptually as we study conglomerate valuations in cross section and empirically as we introduce new methods for a more direct comparison to the external capital market counterfactual).

These groupings are likely to provide improvements in understanding long-run event based returns and for understanding the magnitude of peer effects in corporate finance and asset pricing studies.

Our paper proceeds as follows. In the next section, we describe our data, variables, and methods used to examine product relatedness. We develop new methods to computationally weight peer firms based on their product descriptions and other accounting characteristics. Section III presents validation tests for our replicating peers, and develops an application regarding the valuation of conglomerate and single-segment firms. Section IV presents our methodology and results for how we examine return comovement using text-based replicating peers and compares them to traditional SIC code peers. Section V concludes.

II The Stock Market and Peer Firms

In this section we discuss two hypotheses regarding how the stock market uses fundamental information about peer firms. These hypotheses are rooted in two fundamental channels through which fundamentals impact stock market variables. The first is valuation, and the second is the degree to which stock prices comove over time with peer firms.

Our first hypothesis addresses the link between product uniqueness and firm valuations in the stock market. We separately examine this link for firms that produce in a single industry, and for firms that produce in multiple industries. The literature on product differentiation and uniqueness is extensive. The concept of product differentiation was originally proposed by Chamberlin (1933), with recent articles including Berry, Levinsohn, and Pakes (1997), and Seim (2006))). We focus on stock market valuations and focus more directly on the matter of whether or not firms in the existing universe have adequate product diversity to pose a material replication threat to a given firm under consideration. We postulate that a firm that can be replicated by others firms holds a weaker competitive position and should have a lower valuation as compared to firms that are difficult to replicate. We thus consider the following hypothesis:

H1: Stock Market Valuations and Product Uniqueness: A firm's stock market valuation will be higher when its products are not easily replicated or "spanned" by peer firms (i.e., when the firm is difficult to replicate using best replicating weighted portfolios of rival firms, and hence exhibits product differentiation relative to peer firms).

We measure product uniqueness using the product descriptions of firms and calculating a best portfolio of replicating peer firms whose product descriptions most closely match the focal firm. The specifics of our methodology are described fully in the next section.

Our second hypothesis relates to how firms covary over time in the stock market with their peer firms. This is a joint hypothesis in that it examines both if we have accurately identified peer firms, and if stock market participants also agree regarding the true set of peer firms. To the extent that our identification of peers is in agreement with market participants, firm returns should contemporaneously move together with peer firm returns as they share common fundamentals. For example, this movement will be stronger when these firms share more common supply and demand shocks. This second economic link is tied to our first hypothesis because when firms's products are more unique and difficult to replicate, this comovement should be weaker as more informative peers regarding fundamentals likely do not exist.

A related issue is that the stock market may not contemporaneously recognize economic links. For example, using firms that are linked economically through their suppliers and customers, Cohen and Frazzini (2008) find evidence of ex-ante predictable returns across economically linked firms. They show that the stock market only incorporates return shocks that affect customers with delay. Analogously, we examine if information in peer firm prices is impounded into the price of a given firm immediately, or over time. If the stock market incorporates information about peer firms with a lag, we should also find ex-ante predictable returns over time. Thus we consider both simultaneous stock return comovement (a test of importance of economic links), and whether lagged peer returns predict future own-firm returns (a test of how information is conveyed in the stock market).

H2: Stock Market Comovement and Peer Firms: A firm will comove with its peer firms more when its products are similar to those of peer firms. If the stock market fully recognizes the importance of peer firms, this comovement will be contemporaneous and returns will not be predictable.

III Data and Methodology

A The Sample of 10-Ks

The methodology we use to extract 10-K text follows Hoberg and Phillips (2010a). The first step is to use web crawling and text parsing algorithms to construct a database of business descriptions from 10-K annual filings on the SEC Edgar website from 1996 to 2008. We search the Edgar database for filings that appear as "10-K," "10-K405," "10-KSB," or "10-KSB40." The business descriptions appear as Item 1 or Item 1A in most 10-Ks. The document is then processed using APL to extract the business description text and a company identifier, CIK.[5] Business descriptions are legally required to be accurate, as Item 101 of Regulation S-K requires firms to describe the significant products they offer, and these descriptions must be updated and representative of the current fiscal year of the 10-K.

B Word Vectors and Cosine Similarity

Based on the database of business descriptions, we form word vectors for each firm based on the text in product descriptions of each firm. To construct each firm's word vector, we first omit common words that are used by more than 25% of all firms. Following Hoberg and Phillips (2010a), we further restrict our universe in each year to words that are either nouns or proper nouns.[6] Let M_t denote the number of such words. For a firm i in year t, we define its word vector $W_{i,t}$ as a binary M_t-vector,

[5]We thank the Wharton Research Data Service (WRDS) for providing us with an expanded historical mapping of SEC CIK to COMPUSTAT gvkey, as the base CIK variable in COMPUSTAT only contains the most recent link.

[6]We identify nouns using Webster.com as words that can be used in speech as a noun. We identify proper nouns as words that appear with the first letter capitalized at least 90% of the time in the corpus of all 10-K product descriptions. Previous results available from the authors did not impose this restriction to nouns. These previous results were qualitatively similar.

having the value one for a given element when firm i uses the given word in its year t 10-K business description. We then normalize each firm's word vector to unit length, resulting in the normalized word vector $N_{i,t}$.

Importantly, each firm is represented by a unique vector of length one in an M_t-dimensional space. Therefore, all firms reside on a M_t-dimensional unit sphere, and each firm has a known location. This spatial representation of the product space allows us to construct variables that more richly measure industry topography, for example, to identify other industries that lie between a given pair of industries.

The cosine similarity for any two word vectors $N_{i,t}$ and $N_{j,t}$ is their dot product $\langle N_{i,t} \cdot N_{j,t} \rangle$. Cosine similarities are bounded in the interval [0,+1] when both vectors are normalized to have unit length, and when they do not have negative elements, as will be the case for the quantities we consider here. If two firms have similar products, their dot product will tend towards 1.0 while dissimilarity moves the cosine similarity toward zero. We use the "cosine similarity" method because it is widely used in studies of information processing (see Sebastiani (2002) for a summary of methods). It measures the cosine of the angle between two word vectors on a unit sphere.

C Replicating Peers: Existing Methods

Throughout our discussion of replicating peers, we will adopt the following terminology. We will refer to the firm being replicated as the "focal firm". As we aim to build peer replicas of both congomerate and single segment focal firms, we will generally use the term "pure plays" to refer to the set of replication peers used in a given replication calculation. We use this term for parsimony, and to emphasize that we only consider single segment firms as candidate replication peers, which ensures our measures maintain a transparent interpretation, and that they are not influenced by issues underlying why firms choose to be conglomerates.

Although we depart significantly from the literature, we first consider a modified algorithm based on Lang and Stulz (1994) (LS) and Berger and Ofek (1995) (BO).[7]

[7]Many studies including Campa and Kedia (2002) and Villalonga (2004b) use this methodology.

Although these studies focus on conglomerate focal firms, we note that the methods used by LS and BO apply to single segment focal firms as well, and we consider both types of firms.

LS and BO begin by defining a universe of candidate single segment pure plays to replicate each conglomerate focal firm segment. In BO, this universe is initially defined as all pure plays operating in the firm's four digit SIC industry. However, if the number of pure plays in this universe is less than five, then the pure plays in the given segment's three-digit industry are used. Finally, coarseness is increased to the two digit or even the one digit level until a universe of at least five pure plays is identified. Because changing the level of coarseness can alter the economic information contained in the benchmark (due to economies of scope or irrelevant peers), we exclusively use three-digit SIC industries as our starting point following the broader literature on industry analysis in Finance. However, we can report that using varying levels of coarseness as used in BO does produce materially similar results.

The second step we adopt follows BO's framework, and we compute the firm value to sales ratio for each pure play firm in each focal firm segment's universe (here a focal firm segment could be a single segment focal firm or a segment residing in a conglomerate focal firm), and then we compute the median. The given segment's imputed value is the segment's actual sales multiplied by this median ratio. Medians are used in this literature to reduce the impact of outliers, as firm value to sales ratios can become extreme, especially when firms have low sales or high growth options. Finally, the imputed value of a conglomerate focal firm is the sum of the imputed values of the given conglomerate's segments. For a single segment focal firm, the segment imputed value is the firm imputed value. Excess value is the natural logarithm of the focal firm's imputed firm value divided by the actual firm value. This calculation can also be done using assets as an alternative to sales. A negative excess value, intuitively, suggests that the focal firm is valued less than the value of the peers used to compute the imputed value. We refer to this method as the "Berger+Ofek Baseline" method.

D Replicating Peers: Unconstrained Text-Based Methods

We note three key limitations of the LS and BO methods. A first is the equal treatment of all pure plays in a given segment's pure play universe in the key median calculation. This assumption can reduce accuracy, as additional information exists regarding the nature of the products each pure play produces, and their comparability to a given focal firm being replicated. Methods that weight more relevant pure plays more heavily should perform better. A second limitation is the use of SIC codes to identify the universe of relevant pure play benchmarks. Methods that enhance the set of pure plays beyond traditional SIC boundaries should perform better. A third limitation of the LS and BO method is the focus on a single accounting characteristic such as sales or assets. Candidate pure play firms likely vary along many other dimensions that can also explain valuation differences. For example, some pure plays might have very high sales growth, and might not be relevant as a benchmark for a given mature focal firm. Henceforth, we refer to these three limitations as the "equal weighting limitation", the "limited universe limitation", and the "single characteristic limitation", respectively. Text-based methods offer a solution to all three limitations. In this section, we first examine vocabulary decompositions that directly address the first two limitations. We address the third limitation in the next section.

Although we consider many text-based methods, we adopt the approach of changing one degree of research freedom at a time. Our most basic text-based reconstruction method therefore holds fixed the set of pure-play benchmarks used in BO (those in the same three-digit SIC code). However, we use a textual decomposition to determine weights based on which pure plays use product vocabulary that best matches that of the focal firm. We use these weights to replace the BO equal-weighted median calculation with a weighted median calculation. To determine the weights, we use least squares to decompose the business description of the conglomerate or single segment focal firm being replicated into parts from each of the pure play firms.

Using the same notation from Section III, let M_t denote the number of unique words in the corpus, i denote a given focal firm being reconstructed, t denote the

year of the given focal firm observation, and $N_{i,t}$ denote the focal firm's (M_t x 1) normalized word vector. Further suppose that the given focal firm-year observation has $N_{it,bench}$ candidate benchmark pure play firms to use in its reconstruction. Each pure play has its own normalized word vector. Let $BENCH_{it}$ denote a (M_t x $N_{it,bench}$) matrix in which the normalized word vectors of the benchmark pure plays are appended as columns. We thus identify the set of pure play weights (w_{it}) that best explains the firm's observed product market vocabulary as the solution to the following least squares problem.

$$\underset{w_{it}}{MIN}(N_{it} - BENCH_{it} \cdot w_{it})^2 \tag{1}$$

The solution to this problem (w_{it}) is simply the regression slopes associated with a no-intercept regression of the conglomerate's observed word usage N_{it} on the word usage vectors of the $N_{it,bench}$ pure plays. Importantly, unlike the BO method where pure plays are treated equally, this method assigns greater weight to pure plays whose product vocabulary best matches that of the focal firm. Imputed value is therefore computed by first computing the weighted median value to sales ratio for all $N_{it,bench}$ pure plays using the weights w_{it}. We then multiply the resulting value to sales ratio by the focal firm's total sales to get the conglomerate's imputed value, and excess value is then equal to the natural logarithm of the imputed value to actual focal firm value ratio. We refer to this most basic text reconstruction, which addresses the "equal weighting limitation", as the "SIC Universe: Unconstrained" method.

We next consider an analogous method with a single additional enhancement that also addresses the "limited universe limitation". In this case, we add to the pure play universe by adding pure play firms that are in the focal firm's TNIC industry as defined in Hoberg and Phillips (2010a). These firms have products that are similar to the focal firm's product description, and the TNIC industry classification is equally as coarse as are SIC-3 industries. The calculation follows as described above, except in this case the number of benchmarks $N_{it,bench}$ is as large (if no pure play TNIC peers exist) or larger (if pure play TNIC peers do exist). We refer to this method as the "SIC+TNIC Universe: Unconstrained" method.

E Replicating Peers: Constrained Text-Based Methods

We next consider the third limitation, the "single characteristic limitation". The LS and BO method has an underlying assumption that a single pure play firm characteristic, for example sales or assets, is a sufficient statistic to explain a pure play's firm value. Because asset valuations are forward looking and depend on fundamentals (such as profitability), this limitation can be quite severe. We consider a constrained least squares approach to construct a pure-play based imputed value that holds any number of accounting characteristics fixed to those of the conglomerate itself.

Using the same notation, suppose a focal firm has $N_{it,bench}$ candidate pure play firms. Suppose the researcher identifies N_{char} accounting characteristics they wish to hold fixed when computing imputed valuations. In our case, we consider $N_{char} = 5$, and account for the following five accounting characteristics: Sales Growth, Log Age, OI/Sales, OI/Assets, and R&D/Sales. Let C_{it} denote a N_{char} x 1 vector containing the focal firm's actual characteristics for these five variables. Let Z_{it} denote a $N_{it,bench}$ x N_{char} matrix in which one row contains the value of these five characteristics for one of the pure play benchmark candidates. We then consider the set of weights w_{it} that solve the following constrained optimization:

$$\underset{w_{it}}{MIN}(N_{it} - BENCH_{it} \cdot w_{it})^2 \text{ such that } \; Z'_{it} w_{it} = C_{it} \tag{2}$$

The solution to this problem (w_{it}) is simply the slopes associated with a no-intercept constrained regression of the conglomerate's observed word usage N_{it} on the word usage vectors of the $N_{it,bench}$ pure plays. The closed form solution for the weights is:

$$w_{it} = (BENCH'_{it} BENCH_{it})^{-1}(BENCH'_{it} N_{it} - Z_{it}\lambda), \text{ where} \tag{3}$$

$$\lambda = [Z'_{it}(BENCH'_{it} BENCH_{it})^{-1} Z_{it}]^{-1}[Z'_{it}(BENCH'_{it} BENCH_{it})^{-1} BENCH'_{it} N_{it} - C_{it}]$$

Intuitively, this set of weights identifies the set of pure plays that use vocabulary that can best reconstruct the focal firm's own vocabulary, and that also exactly match the focal firm on the N_{char} characteristics. We refer to this method as the "SIC+TNIC Universe: Constrained" method.

F Replicating Peers: Accounting for Segment Sales

The discussion in this section is only relevant for conglomerate firms, and the objective is to potentially account for the fact that the conglomerate segment tapes not only contain information about the industries in which a conglomerate operates, but also information about how large each segment is. This data is in the form of sales, which are reported at the segment level.

The LS and BO method computes imputed values segment-by-segment, and therefore utilizes information contained in reported segment-by-segment sales. To the extent that sales explains valuations better than other characteristics, this information might be useful. The basic text-based methods described above do not use segment-by-segment sales, and instead rely on the weights obtained from the textual reconstruction to derive imputed value. We believe that it is an empirical question as to whether textual weights or sales weights best explain valuations. However, it is important to explore this question. We therefore consider a method that is identical to the "SIC+TNIC Universe: Constrained" method described above, except that we add an additional set of constraints based on the segment sales to ensure that the imputed value is weighted by sales across segments as is the case for the BO method.

Consider a conglomerate focal firm having $N_{it,seg}$ segments, and let S_{it} denote the $N_{it,seg}$ x 1 vector of sales weights (one element being a given segment's sales divided by the total sales of the conglomerate). To compute imputed values that impose segment sales-based weights, we make two modifications to the constrained optimization. First, we append the vector S_{it} to the vector C_{it}. Second, we create a $N_{it,bench}$ x $N_{it,seg}$ matrix of ones and zeros. A given element is one if the pure play associated with the given row is in the industry space corresponding to the given segment of the conglomerate focal firm associated with the given column. This matrix is populated based on how the pure-play benchmarks are selected. If the benchmark is selected due to its residing in a three digit SIC industry of a given segment, then the given pure play firm is allocated to that segment. If the benchmark was selected due to its residing in the TNIC industry of the conglomerate focal firm itself, then it is allocated to the segment whose SIC-benchmarks it is most similar

(as measured using the cosine similarity method). We then append this $N_{it,bench}$ x $N_{it,seg}$ matrix of ones and zeros to the matrix Z_{it}. The solution to the resulting constrained optimization is a set of new weights w_{it} that has the property that the sum of weights allocated to each segment equals the given segment's sales divided by the total conglomerate sales ratio. Therefore, imputed values can be computed segment by segment for the focal conglomerate firm. We refer to this method as the "SIC+TNIC Universe: Constrained, Segment-by-Segment" method.

IV Results: Firm Valuation

In this section, we first assess the similarity of replicating peers using the reconstruction methods discussed in the previous section. Because the literature on replicating peers has a long history of focusing on conglomerates, and because the reconstruction methods materially differ for conglomerates and single segment firms, we separately present results for conglomerates and single segment firms. This separate reporting also ensures the comparability of our results regarding past studies. In particular, many past studies in the conglomerate literature consider benchmarking in the analysis of conglomerate excess valuations, and hence the results in our study can thus be more directly compared to the methods used in those studies.

After we assess the similarity of replicating peers, we briefly readdress the question of whether or not conglomerates trade at a discount. We then conclude this section by testing a hypothesis regarding whether the existence of high similarity replicating peers explains firm valuations in cross section. In particular, we extend the literature and we examine this hypothesis for both conglomerate and single segment firms.

A Methodological Validation

Following the methodology discussion in Section III, we examine excess valuations using five different replication peer identification methods. In particular, we consider five methods discussed in the previous section for identifying replicating peers: the Berger and Ofek (1995) benchmark, and four text-based methods aimed at addressing key limitations in the BO method. Table I (conglomerates) and Table II (single

segment firms) display average excess valuations, and mean squared error statistics based on these five methods.

As discussed earlier, excess value calculations have been examined extensively in the conglomerate literature, and early studies find that conglomerates trade at discounts by illustrating that average excess valuations are negative. More recent studies suggest that this result result is not robust to various enhancements including selectivity controls. We examine average excess valuations for comparability with existing studies, and we also consider mean squared error (MSE) statistics to compare relative valuation accuracy across valuation methods. A method with a lower MSE generates predicted valuations that are closer to actual valuations, and is therefore more accurate.

Panel A of each table presents summary statistics based on raw data. Following convention in the literature in Panel B of each table, we discard an excess value calculation if it is outside the range $\{-1.386, +1.386\}$ (in actual levels instead of natural logs this range is $\{\frac{1}{4}, 4\}$), to reduce the effect of outliers. Therefore, the observation counts available for each valuation method vary slightly as more accurate valuation methods generate excess valuations outside this range less often, and thus have higher observation counts. In Panel C, we omit a firm-year for all calculations if its estimated excess value is outside this range using any calculation method we consider (as this allows a comparison that holds the sample size fixed across all methods). Both tables report mean excess value, MSE statistics, and observation counts for excess value calculations based on sales (first three columns) and assets (last three columns).

Following conventions in the literature, we apply additional screens to the sample included in this part of our study. In particular, we require lagged COMPUSTAT data, we drop firms with sales less than \$20 million, firms with zero assets, we require that 10-K text data is available, and also that a sufficient number of pure play firms exist in segment industries to compute excess valuations. Regarding comglomerates, we apply one additional screen following existing studies, and we discard conglomerates for which summed segment sales disagrees with the overall firm's sales by more than 1%.

[Insert Table I Here]

Table I displays results for conglomerates. Panel A shows that as more refined text-based valuation methods are used, the conglomerate discount disappears. For excess valuations based on sales, the 8.2% discount for the Berger and Ofek benchmark in row one declines to just 1.2% using the text-based method that addresses all three limitations. The most basic text-based benchmark, which holds fixed the same SIC-universe of pure play candidates, results in a decline in the excess value discount to 5.8%. Therefore, changing the weighting of single segment firms from equal weighting as in Berger and Ofek to textual importance weights is partially but not fully responsible for our ability to explain the discount. Row 3 of Panel A expands the universe of firms eligible to receive positive weights to include the TNIC pure play rivals of the conglomerate. This expansion further reduces the discount to 4.6%. Finally, matching jointly on both the textual vocabulary dimension, and the five key accounting characteristics, Row 4 shows that the discount declines nearly to zero at 1.2%. In row 5 of Panel A, we see that further constraining the weights to match segment-specific sales ratios has little relevance as the discount changes little to 1.8%.

When excess valuation is based on assets in the fourth column, we see that the discount of -2.7% using the Berger and Ofek benchmark declines analogously to nearly zero (0.1%) using the constrained text-based benchmark in row four. We conclude that our ability to explain the benchmark is due to three factors: (1) Using weights based on textual decompositions, (2) improving the benchmark candidates to include both SIC and TNIC peers, and (3) constraining the benchmark to have similar accounting characteristics relative to the conglomerate being reconstructed.

Columns two and four, which report mean squared error statistics, strongly support the conclusion that the constrained model based on the enlarged SIC+TNIC universe offers the most accurate set of conglomerate replicating peers. When based on sales, the mean squared error in row 4 of 0.320 reaches a minimum and is 32.4% smaller than the mean squared error of 0.474 associated with the Berger and Ofek benchmark. When based on assets, this improvement is almost as large at 27.7%.

These results suggest that the improvements in accuracy are very large in economic terms.

In Panels B and C, we omit excess valuations outside the interval $\{-1.386, +1.386\}$. Panel C omits the firm-year observation if any of the five valuation method places the value outside this range. The results are similar to Panel A. We see the excess value discount disappearing using our text-based methods, and we also observe mean squared error decreasing. In Panel C, the excess value discount entirely disappears for both the sales based and the asset based methods. The results in Panel C are especially clean because the sample size is held fixed across methods.

We conclude that using higher similarity replicating peers can fully explain the previously reported conglomerate discount, and also dramatically improve valuation accuracy. The intuition behind this result squares well with the original intent: a portfolio of pure plays that matches the conglomerate in operations and assets should be a valid benchmark for the conglomerate itself, and it represents a more accurate benchmark regarding how the conglomerate would be valued if it instead operated its segments as a portfolio of single segment firms. Our results therefore support recent studies and do not support the conclusion that conglomerate firms trade at discounts. Other recent studies that draw the same conclusion using other methods include Campa and Kedia (2002), Villalonga (2004b), and Graham, Lemmon, and Wolf (2002). We view this result as important to illustrate that our methods are well constructed and consistent with existing studies. However, we do not view our conclusion that the conglomerate discount vanishes to be a central result given these existing studies.

[Insert Table II Here]

We now turn attention to single segment firms, which are not examined in the aforementioned literature. Interestingly, using the Berger and Ofek (BO) method, we also see an initial discount of nearly 5% for single segment firms in Panel A. As it does for conglomerates, this discount also disappears when text based valuation methods are used. Furthermore, Panel B and Panel C show that the discount also disappears when outliers are dropped, both for the BO method and for text-based

methods. That is, regardless of how outliers are handled, the excess values remain close to zero for text based replicating peer methods.

More importantly, the MSE calculations in Table II affirm that text based methods also offer substantial improvements in valuation accuracy for single segment firms. Moreover, the more sophisticated text-based methods offer the best improvements of all. For example, MSE declines from 60.2% for the BO method using raw data in Panel A to 50.7% using expanded TNIC and SIC peers, and then declines further to 41.7% when we further construct replicating peers based on both text and accounting data. These dramatic improvements are remarkably similar to the improvements noted above for conglomerates. For example, both single segment firsm and conglomerate firms experience an improvement in MSE of roughly 32% in comparing the BO benchmark to the "SIC+TNIC universe: Constrained" model.

We view these comparisons across single segment and conglomerate firms to be critical regarding the directions taken in the remainder of the paper. We continue to report many key results for both conglomerate and single segment firms. However, we will later draw the conclusion that text-based benchmarks, as they are driven by the uniform principal that even a complex firm's product offerings can be reconstructed using a large group of replicating peers, offer advantages in benchmarking that are universal to both conglomerate and single segment firms.

B Replicating Peers and Accounting Characteristics

In Table III, we assess whether replicating peers have similar average accounting characteristics as the conglomerate (Panel A) or single segment firms (Panel B) they intend to replicate. As the objective of these methods is to rebuild an identical replica of any given firm, better peers should match the focal firm along many dimensions beyond valuation (discussed above). For example, they should have similar sales growth, they should be equally as mature, as profitable, and they should have similar investment intensities.

To assess this prediction, we first compute the implied characteristics for each accounting variable using the same methods used to compute imputed valuations in

the excess valuation calculation. For example, the implied Sales Growth of a Berger and Ofek (baseline) benchmark is computed as the equal weighted median of the given characteristic for pure play firms in the same three digit SIC code as the given segment. For textual methods, we simply use the weighted median sales growth using the same set of textual weights as before. This calculation is analogous for single segment and conglomerate firms, as each simply implies a different set of weights as discussed in the methods section.

[Insert Table III Here]

Table III reports correlations between the actual firm characteristics and the implied replicating peer characteristics for each characteristic noted in the first column using each replication method noted in the remaining columns. Higher correlations indicate that the replication was more successful in matching the true firm for the given characteristic. For conglomerate firms, Panel A reveals that the text-based benchmarks strongly outperform the Berger and Ofek benchmark for every single characteristic assessed. Even the simplest text-based methods (that do not constrain accounting characteristics) in columns two and three have significant improvements in correlations compared to the BO correlations in the first column. For example, the 28.9% correlation between the OI/Assets for the BO benchmark increases dramatically to (35.7% to 42.1%) using these simple unconstrained text-based weights.

As indicated in the methodology section, the unconstrained text-based weights are purely a function of the vocabulary used by the pure plays and the focal firm, and are not mechanistically related to the accounting numbers that these methods better match in these tests. In the last two columns, not surprisingly, we observed that Pearson correlations rise dramatically when we use the text-based constrained optimization. As these weights constrain the replicating peers to match the focal firm on five key accounting characteristics, it is thus not surprising that these characteristic correlations are dramatically higher. We conclude that text-based measures offer substantial improvements over existing methods.

Panel B of Table III shows that improvements in these correlations also exist for single segment firms, but also that the improvements are less dramatic for the

simplest text based methodology. Some correlations decline slightly from the BO benchmark to the SIC-only unconstrained textual method in the second column. For example, the sales correlation dips from 20.4% to 18.1%, indicating that equally weighting peers can match somewhat better in terms of size (although the text based methods uniformly outperform on key fundamentals including profitability, investment style, and Tobin's Q). Despite this rather modest result for the simplest text based replication peers, the later columns illustrate that more elaborate text based peers outperform BO benchmarks on all characteristics, and by a large margin.

It is also natural to ask which type of replicating peers are weighted more than others when reconstructing conglomerates and single segment firms. Panel A of Table IV explores this question for conglomerates, and Panel B for single segment firms. Both panels display average accounting characteristics for the replicating peers that are assigned high weights (those in the highest quartile using the text-based conglomerate benchmarks) versus those assigned low weights (those in the lowest quartile). In particular, we construct a large database of high weighted replicating peers based on sorting the peers from each focal firm-year replication into quartiles, and extracting those in the high quartile. We build a similar database for low weighted peers, and we formally compare characteristics across the two databases to examine systematic differences in the firms receiving high versus low weights. This test is not possible using the historical Berger and Ofek method, as that method assigns equal weights to all firms. Our framework generates a very powerful test of peer attributes, and can shed new light on issues underlying peer selection for conglomerate firms versus single segment firms.

The first three columns of Panel A are based on the "SIC+TNIC universe (unconstrained)" method. This method is text-based and uses an enhanced set of eligible replicating peers (SIC and TNIC peers) to reconstruct a given conglomerate. In the second three columns in Table IV, we repeat the same exercise using the "SIC+TNIC universe (constrained)" method, which also holds fixed key accounting variables when identifying replicating peers as discussed earlier.

[Insert Table IV Here]

Panel A shows shows that pure play firms receiving higher weights using text decompositions tend to be older, are more mature firms, and have lower sales growth. These firms also have less research and development, and are more profitable than those pure plays assigned lower weights. Because mature firms have lower valuation ratios, this helps to explain why conglomerates appear undervalued using earlier methods.

The results in the latter three columns are similar to those in the first three columns, but are notably sharper. For example, the average difference in age is nearly 7.5 years using the constrained text method, compared to just 4.4 years using the unconstrained text method. We conclude that equal weighting all pure plays, as was done using the Berger and Ofek benchmark, will overweight high growth firms and thus generate the unwarranted conclusion that conglomerates are undervalued. Our results in the next section formally confirm this conjecture.

Panel B shows that similar results do not obtain for single segment replicating peers. This result should not be surprising given that all decompositions are based only on single segment firms to maintain a clear interpretation and to maintain consistency with earlier literature. More succinctly, replicating peers, which are limited to single segment firms, are unlikely to be systematically different from the single segment firms they aim to replicate. We thus do not observe material differences in the firms receiving high versus low weights regarding their size and profitability, and more generally significance levels and difference magnitudes are substantially smaller in Panel B for single segment firms when compared to the conglomerate firms in Panel A.

C Stock Market Valuation and Difficulty to Replicate

In this section, we examine a key hypothesis relating to the economics of replication peers can explain the extent to which firm valuations vary in cross section. We hypothesize that focal firms that are harder to replicate using replication peers will have higher valuations relative to firms that are more easily replicated. In particular, firms that are harder to replicate using their product text descriptions are likely to

have more unique products, and likely face less direct product market competition as well as less severe competitive threats. A firm that is easy to replicate might face pressure in the product market if its product market exhibits high profitability. For example, the existence of nearly perfect replicating peers indicates that the replicating peers themselves can likely enter the given focal firm's market at low cost. Firms that are difficult to replicate cannot be credibly pressured through this same channel, as high similarity peers do not exist.

To explore this question, we regress both conglomerate and single segment firm excess valuations on the measure of difficulty to replicate generated by the text based replication peers optimization problem. In particular, this variable is one minus the R^2 associated with the textual decomposition, which is quantitatively analogous to a constrained regression model. Because the existing literature focuses extensively on the excess valuation of conglomerate firms, we separately examine our key hypothesis within the sample of conglomerate firms and single segment firms. We also include controls for document length, and accounting variables used in the existing literature.

[Insert Table V Here]

Table V displays the results of OLS panel data regressions in which one observation is one conglomerate focal firm in one year (Panels A and B), or one single segment focal firm in one year (Panels C and D). The dependent variable is the focal firm's excess valuation using the constrained text-based valuation method (Panels A and C) and the Berger and Ofek (1995) valuation method (Panels B and D). t-statistics are shown in parentheses, and standard errors are adjusted for clustering by firm. We also standardize all independent variables to have a standard deviation of one for ease of interpretation and comparison of coefficients.

Our first key finding is that the difficulty of pure plays to replicate variable - both for other single-segment firms and for conglomerate firms - is positive and highly statistically significant in all four panels. Because the independent variables are standardized, we can also interpret the coefficients to mean that a one standard deviation increase in difficulty to replicate generates a 5% to 9% increase in valuation. Both conglomerates and single segment firms that are harder to replicate have

higher valuations relative to their replication peers. As this variable captures the uniqueness of the conglomerate's products relative to its best replicating peers, one would not expect its affect on valuation to be negated out in the difference used to compute excess valuations. Unlike many variables, which have industry and firm level components, this variable is a unique property of any given focal firm that is not necessarily a property of its its industry peers.

Our findings regarding the difficulty to replicate, which are robust at the 1% level of significance in all specifications, are consistent with unique firms earning higher rents due to the inability of other firms to replicate their unique products. The consistent results of similar magnitude for conglomerates and single segment firms alike also indicates that uniqueness rents can likely be generated in many forms. For example, barriers to entry can create a scenario in which a single segment firm can achieve a high degree of difficulty to replicate. Conglomerate firms can generate gains through this same channel, or the conglomerate structure itself can be used to assemble divisions that, when combined, are difficult to replicate by virtue of product market synergies that require multiple technologies from the multiple segments. Our control variables indicate that firms are also valued more when they have more investment (R+D and Capital Expenditures), when they are more profitable, and when they are larger.

We also find that the reported R^2s are higher in Panels B and D compared to those in Panels A and C. This result arises because our text-based valuation methods produce benchmarks that are more comparable to the given conglomerate (as shown previously). Hence, spurious differences in valuation relating to mismatched characteristics are less likely using text based methods as Panels A and C illustrate. Put differently, excess valuations are more difficult to predict or explain when systematic biases in measurement are removed. The table also shows that the level of significance of our key variable, difficulty of pure plays to replicate, is quite similar in all panels, and thus it is robust to changes in the the replicating peer methodology, as well across conglomerate and single segment firms.

We next assess the economic magnitudes of our findings regarding the difficulty of pure plays to replicate variable. In each year, we sort firms into quintiles based on

this variable, and we compute the average excess valuation for each group. We also compute the average residual excess valuation, where residuals are from a regression of excess valuation on all of the variables in Table V with the exception of the difficulty to replicate variable. We compute results separately for conglomerate and single segment firms.

[Insert Table VI Here]

Table VI displays the results for conglomerate firms in Panel A and single segment firms in Panel B. Panel A shows that raw conglomerate excess valuations are modestly higher for the highest quintile (+5.4% using the text-based model) relative to the lowest quintile (-2.3%). This effect is magnified for average residual excess valuations (+9.1% versus -4.8%). Panel B shows an analogous result for single segment firms. The modest inter-quintile range of nearly 4% for raw excess valuations increases to nearly 14% for residual excess valuations. We conclude that the impact of a firm's difficulty to reconstruct using replication peers is meaningful, and that both conglomerates and single segment firms that are more difficult to replicate trade at modest premia relative to their replication peer benchmarks.

Overall, our results are consistent with firms having higher valuations when their products are difficult to replicate. This suggests that such firms extract more value through differentiated product offerings that cannot be easily raided by potential rival peer firms. Going further, in unreported tests, we find that the difficulty to replicate does not predict abnormal stock returns. Hence, the high valuations associated with firms having higher difficulties to replicate are likely permanent, and reflect the stock market recognizing the value it entails based on firm fundamentals such as protection from rivals.

V Stock Market Peer Comovement and Product Uniqueness

In this section, we examine if focal firm monthly stock returns comove more with benchmark portfolios constructed from replicating peers versus those constructed

from historical three digit SIC-based portfolios. The purpose of this test is threefold. First, as stock returns are driven at least in part by fundamentals, this test serves as a benchmark regarding how researchers can assess the similarity of text based replicating peers versus SIC-based peers (as used in Berger and Ofek and subsequent studies to examine valuation). This test of benchmark quality is particularly clean, as unlike valuation which is a cumulative statement of value, monthly stock returns reflect high frequency changes in valuation. Hence, these results are unlikely to be influenced by issues underlying long-lived drivers of accounting or valuation outcomes. Second, whereas our previous examinations are more linked to studies in corporate finance, this test examines the relevance of our replicating peers for applications in asset pricing, where the focus has been primarily on stock returns rather than on valuations. Third, this framework allows us to evaluate the informational efficiency of the stock market by examining whether information in peer stock returns is impounded in the stock market immediately, or with a material lag. Finally, this framework allows us to examine one additional prediction underlying the issue of a firm's difficulty to replicate, as examined in earlier sections. In particular, if replicating peers are indeed generated using product market fundamentals, it should be the case that stock market comovement signals are stronger when a firm is easy to replicate and hence the replicating peers as candidate benchmarks are superior.

Because other metrics used to generate replicating peers including historical SIC codes to not provide data structures indicating the quality of the replication for each firm in a given group, our study is the first study to our knowledge to explore the role of peer similarity in driving stock market comovement.

A Stock Market Comovement

We first explore the unconditional link between the return of a given focal firm and a weighted portfolio of replicating peers. We thus consider the returns of three feasible investment portfolios, and we then explore the comovement properties of these portfolios. The first investment is simply a long position in a single focal firm's stock, where the focal firm can be a conglomerate or a single segment firm (we examine both separately and we consider sample-wide results). The second portfolio

is a long position in an equal weighted portfolio of firms residing in the same three-digit SIC code as the given focal firm. We will refer to the returns of this portfolio as the "SIC-3 Peer Return". The third position is a long position in the weighted portfolio of replication peers constructed using the best text-based methodology, which is the "SIC+TNIC universe (constrained)" method, which also holds fixed key accounting variables when identifying replicating peers as discussed earlier. This portfolio utilizes the same weights used in all earlier tests in this paper, as they sum to one, and this portfolio produces a counterfactual to the given focal firm that best replicates both its product offerings and its accounting characteristics. We refer to this portfolio's return as the "Text-based Peer Return".

We then consider Fama and MacBeth (1973) regressions in which the focal firm's monthly stock return is the dependent variable. One observation is one firm in one month from 1997 to 2008. The independent variables include the SIC-3 Peer Return (excluding the firm itself) and the text-based peer return (also excluding the firm itself), both measured using stock return data from month t. We also include controls for the Fama and French (1992) variables (log book to market ratio and log size), a dummy for negative book to market ratio stocks, and a control for momentum (defined as the own-firm 11 month lagged return from month $t-12$ to $t-2$).[8]. We consider focal firm stock returns from month t to examine simultaneous comovement, and focal firm stock returns from month $t+1$ and $t+2$ to examine the extent to which each variable predicts returns, as would be predicted under a hypothesis of lagged information dissemination.

[Insert Table VII Here]

Table VII displays the results, and Panel A displays results for all firms, and Panels B and C display results for conglomerates and single segment firms, respectively. Panel D displays results for all firms, however, we consider various lags to

[8]We do not display the negative book to market dummy to conserve space and because it is not statistically significant. We include the dummy so that we can set the log book to market ratio to zero for these observations and include them in the regression. Following convention, our momentum variable is based on 11 months and excludes the most recent month to avoid contamination with microstructure issues. Our results are not sensitive to dropping the negative book to market observations or to defining momentum to include the most recent month.

the dependent variable as noted in the dependent variable column. All independent variables are standardized to have a standard deviation of one for ease of comparison.

Panel A shows that the text based peer return is substantially more informative than the SIC-3 peer return in explaining simultaneous comovement with a given focal firm's monthly stock return. This results affirms that text based replicating peers likely better reflect a given focal firm's fundamentals, and hence its stock returns. Because the two right hand side variables are standardized, the coefficients can also be compared, and this reveals that the extent to which text-based peers are more informative is very large. In simultaneous regression, the text based peers are roughly five times as informative, and in univariate regressions, it is nearly twice as large. Overall text based peers are substantially more informative, but SIC-3 peers still retain a modest amount of unique information after controlling for text-based peers.

Panels B and C show that text-based peers are equally as informative in explaining both conglomerate and single segment firm stock returns. However, we observe substantial degradation in the ability of SIC-3 peers to explain conglomerate returns. This is likely due to the fact that the error in SIC codes is compounded when a researcher must rely on SIC codes not only for a firm's overall product offerings, but must also rely on the segment tapes and their corresponding SIC codes to construct a materially more complex firm. The text-based peers are directly generated from the given focal firm's unique product market vocabulary from its 10-K, and thus do not rely on the accuracy of the Compustat segment tapes.

Panel D examines the extent to which lagged peer returns predict own-firm monthly stock returns. We first repeat the result from row 1 (simultaneous return) for comparison, and then present results for a one month lag and a two month lag. The table shows that text-based peers are overwhelmingly more successful in predicting ex-post stock returns, and that the results are economically large. As stated above, the right hand side variables are standardized and hence the coefficients can be compared, as well as intuitively interpreted. The results suggest that both peer variables predict next-month returns, and that a one standard deviation higher text-based peer return generates a 1.2% improvement in own firm stock return, whereas a one standard deviation higher SIC-3 peer return only generates a 0.3% improvement

in own firm stock return. Finally, considering a two month lag, we observe that SIC-3 peers do not predict own firm stock returns in a significant fashion, whereas a one standard deviation higher text based peer return generates a still-significant 0.5% improvement in own-firm stock return. These results are economically important in size and compare well with other variables used to predict stock returns in the anomalies literature. These findings are consistent with information associated with text-based peers disseminating more slowly than information associated with SIC-3 based peers.

B Comovement and Product Uniqueness

In this section, we examine the prediction that firms with more similar products - and thus better replicating peers - experience stock market comovement that is more strongly related to the returns of its text-based peer portfolio. We also test whether our previous findings regarding return predictability are also linked to the similarity of replicating peers, as a stronger signal should generate larger predictable returns under the assumption that information disseminates slowly, and that the signal dissipating in any particular time interval is monotonically related to the magnitude of the link to firm fundamentals.

[Insert Table VIII Here]

Table VIII performs this analysis, and as before, we consider Fama-MacBeth regressions in which the own-firm monthly stock return is the dependent variable. However, in this case, we consider conditional results, where we condition on high similarity replicating peers (firms that are "easy" to replciate using text based peers) and on low-similarity replicating peers. These low-similarity firms are difficult to replicate as they have more unique products. Panel A presents results for all firms, and the specifications match those in Panel D of Table VII. Panel B presents analogous results for firms that are difficult to replicate, ie, those with above median difficulty to replicate (one minus the R^2 from the best textual replicating peer identification model, which is the "SIC+TNIC universe (constrained)" method). Panel C presents results for firms with below median difficulty to replicate.

The table shows a strong link between return comovement, return predictability, and the similarity of replicating peers. In particular, all results in Panel C for the text based peer return variable are stronger than those in Panel B, and hence comovement is meaningfully stronger when peers are of higher similarity. As the right hand side variables are standardized, the results are easily interpreted in magnitudes. A one standard deviation upward shift in the text-based peer return, when a focal firm has below median difficulty to replicate (more similar products), generates a 1.8% improvement in predictable own-firm monthly stock returns at a one month lag, and 0.6% improvement at a two month lag. In Panel B, these figures are less than half as large at 0.8% and 0.3%. The table also shows that the impact of simultaneous returns is also roughly twice as large when the peers are of higher similarity and thus the firm has more similar products. Analogously, we do not observe meaningful differences in the link to SIC-3 based peers. This result is expected as the similarity of peers metric is a statement about the similarity of text-based peers, and there is no analogous metric for SIC-3 based peer similarity.

We conclude that the difficulty to replicate variable is not only relevant in a corporate finance setting, as it strongly relates to conglomerate excess valuations, but also in an asset pricing setting, as it identifies firms for which a strong benchmark portfolio exists, while also pointing out stronger return predictability related to delayed dissemination of information.

C Idiosyncratic and Systematic Risk

In this section, we examine the implications of our findings through the lens of systematic and idiosyncratic risk. Because systematic risk is by nature pervasive among large groups of firms, we expect higher degrees of simultaneous comovement from systematic risk shared among peers. Furthermore, if asset prices are more efficient in accounting for systematic price changes, we should also observe less delay in its dissemination and hence lagged peer returns should impact the given firm's current return less. Conversely, if the market is less efficient regarding idiosyncratic return spillovers, perhaps because they are unique and more difficult to value, their impact will be impounded into returns with a longer delay.

To examine this hypothesis, we once again consider Fama-MacBeth regressions with own-firm monthly stock return as the dependent variable. The independent variables include the systematic and idiosyncratic portions of the text-based return benchmark. To compute the systematic portion, we first regress (for each month) daily stock returns for each firm onto the three Fama French factors and the momentum factor. The projection from this regression (excluding the projection from the intercept) is the systematic portion of a firm's daily return. These daily systematic returns are then aggregated to monthly observations, and we compute the average of these monthly systematic returns over each firm's text based peers to compute the "Systematic Peer Return". The "idiosyncratic Peer Return" is then simply the raw text-based peer return (discussed in the previous section) minus the systematic peer return.

[Insert Table IX Here]

The results are presented in Table IX, and Panel A displays results for all firms. Importantly, the systematic and idiosyncratic return variables are standardized to have a mean of zero and a standard deviation of one, and hence their coefficients can be compared regarding importance and impact on the dependent variable. Row (1) of the table shows that a larger fraction of the simultaneous return comovement is attributable to systematic risk. A one standard deviation shift of the systematic peer variable generates a return spillover of 5.5%, compared to just 3.3% for the idiosyncratic variable. In contrast, as we lag the spillover variable, the idiosyncratic variable grows in relative importance. For example, by the third lag in row (4), the idiosyncratic variable still exerts spillovers onto the focal firm's return, whereas the the systematic variable loses both statistical and economic relevance.

The magnitude of the idiosyncratic spillover at 0.3% per month may seem small relative to the magnitude of the simultaneous spillover. However, the lagged effects are predictive, whereas the simultaneous result is not. These results support the conclusion that systematic risk exerts more comovement onto peer returns, however, idiosyncratic returns are subject to more delay in the dissemination of information. Thus the idiosyncratic returns still contain stock market relevant information.

Panels B to E of Table IX display results for various subsamples. Comparing Panel B, above median difficulty to replicate, with Panel C, below median difficulty to replicate, suggests that the magnitude of both systematic and idiosyncratic peer returns are larger for firms that are easier to replicate using peers. Furthermore, we continue to observe a much more rapid dissemination of systematic information relative to idiosyncratic information. Comparing conglomerate firms and single segment firms in Panels D and E, respectively, our results suggest that these same conclusions are uniformly true for both firm types, despite conglomerate firms requiring more intricate replication using product market variables as discussed earlier.

VI Conclusions

We examine how the stock market uses information about firm product offerings and peer firms using text-based computational methods. We show that the stock market incorporates firm product and accounting information of peer firms in firm cross-sectional valuation and in return comovement. Our peer firms are identified using text-based analysis that assigns replicating peers different weights based on how each uniquely contributes to creating a near replica (on the basis of both product offerings and accounting characteristics) of a given firm under consideration. We provide a closed form solution that identifies the best set of replicating peers for any conglomerate or single segment firm. This calculation also generates a new measure of peer similarity based on the extent to which the replica's vocabulary in fact matches the firm being replicated.

We find that firms whose products and characteristics are more difficult to replicate using peer firms have higher stock market valuation and more stock market comovement in returns. These results hold for both both conglomerate and single segment firms, consistent with firms with more unique products being valued by investors and with these firms being able to maintain a competitive advantage. These higher valuations are based on fundamentals and are long-lasting, as we do not observe ex-post stock return reversals.

We show that both contemporaneous peer returns and lagged peer returns predict

own firm stock returns. We document that text-based replication peers dramatically outperform SIC-based peers along both dimensions. On a simultaneous basis, text-based replication peers comove substantially more than SIC-based peers. Moreover, lagged returns based on text-based replicating peers outperform SIC-based peers in predicting ex-post own-firm returns. These results are economically large in magnitude, as comovement is as much as five times larger using text-based replicating peers. We find that comovement links are stronger when firms are more similar and easier to replicate with text-based replication of their product text using peer firms. Lastly, decomposing of peer returns into systematic and idiosyncratic components, we find that simultaneous comovement is most linked to systematic returns, and lagged information is most relevant for idiosyncratic returns. These results are consistent with the market being less informationally efficient regarding idiosyncratic information, which may be unique and more difficult to value.

Our results broadly show that the stock market values product uniqueness. Our results are consistent with peer firms with similar products facing similar cost and demand shocks and thus having similar stock market valuations and return comovement. Firms with unique products that are difficult to replicate with peer firms have higher stock market valuations and less stock market comovement with their peers. Overall, our methodology and new peer groupings allow more accurate benchmarking of competitor firms that should be useful in other corporate finance or asset pricing questions where performance benchmarking or counterfactual analysis is important.

References

Berger, Phillip, and Eli Ofek, 1995, Diversification's effect on firm value, *Journal of Financial Economics* 37, 39–65.

Berry, Steven, James Levinsohn, and Ariel Pakes, 1997, Automobile prices in market equilibrium, *Econometrica* 63, 841–890.

Campa, Jose, and Simi Kedia, 2002, Explaining the diversification discount, *Journal of Finance* 57, 1731–1762.

Chamberlin, EH, 1933, *The Theory of Monopolistic Competition* (Harvard University Press: Cambridge).

Chevalier, Judith A., 1995, Capital structure and product-market competition: Empirical evidence from the supermarket industry, *American Economic Review* 85, 415–435.

Cohen, Lauren, and Andrea Frazzini, 2008, Economic links and predictable returns, *Journal of Finance* 63, 1977–2011.

Comment, Robert, and Gregg Jarrell, 1995, Corporate focus and stock returns, *Journal of Financial Economics* 37, 61–87.

Custodio, Claudia, 2010, Mergers and acquisitions accounting can explain the diversification discount, Arizona State University Working Paper.

Fama, Eugene, and Kenneth French, 1992, The cross section of expected stock returns, *Journal of Finance* 47, 427–465.

Fama, Eugene, and J. MacBeth, 1973, Risk, return and equilibrium: Empirical tests, *Journal of Political Economy* 71, 607–636.

Graham, John, Michael Lemmon, and Jack Wolf, 2002, Does corporate diversification destroy value?, *Journal of Finance* 57, 695–720.

Hoberg, Gerard, and Gordon Phillips, 2010a, Text-based network industry classifications and endogenous product differentiation, University of Maryland Working Paper.

Khanna, Naveen, and Sheri Tice, 2000, Strategic responses of incumbents to new entry: The effect of ownership structure, capital structure and focus, *Review of Financial Studies* 13, 749–779.

Lamont, Owen, and Christopher Polk, 2002, Does diversification destroy value? evidence from the industry shocks, *Journal of Financial Economics* 63, 51–77.

Lang, Larry, and Rene Stulz, 1994, Tobin's q, corporate diversification, and firm performance, *Journal of Political Economy* 102, 1248–1280.

Leary, Mark, and Michael Roberts, 2010, Do peer firms affect corporate capital structure?, Working Paper, University of Pennsylvania.

Lins, Karl, and Henri Servaes, 1999, International evidence on the value of corporate diversification, *Journal of Finance* 54, 2215–2240.

MacKay, Peter, and Gordon M. Phillips, 2005, How does industry affect firm financial structure?, *Review of Financial Studies* 18, 1433–66.

Maksimovic, Vojislav, and Gordon Phillips, 2007, *Conglomerate Firms and Internal Capital Markets, Handbook of Corporate Finance: Empirical Corporate Finance* (North-Holland).

Moskowitz, Tobias J., and Mark Grinblatt, 1999, Do industries explan momentum?, *Journal of Finance* 54, 1249–1290.

Phillips, Gordon M., 1995, Increased debt and industry product markets: An empirical analysis, *Journal of Financial Economics* 37, 189–238.

Rauh, Joshua, and Amir Sufi, 2010, Explaining corporate capital structure: Product markets, leases, and asset similarity, Northwestern University Working Paper.

Sebastiani, Fabrizio, 2002, Machine learning in automated text categorization, *ACMCS* 34, 1–47.

Seim, Katja, 2006, An empirical model of firm entry with endogenous product choices, *Rand Journal of Economics* 37, 619–40.

Servaes, Henri, 1996, The value of diversification during the conglomerate merger wave, *Journal of Finance* 51, 1201–1225.

Stein, Jeremy, 2003, Agency, information and corporate investment, *Handbook of the Economics of Finance* pp. 110–63.

Villalonga, Belen, 2004a, Diversification discount or premium? new evidence from business information tracking series, *Journal of Finance* 59, 479–506.

——— , 2004b, Does diversification cause the diversification discount, *Financial Management* 33, 5–27.

Wernerfelt, Birger, and Cynthia Montgomery, 1988, Diversification, ricardian rents, and tobin's q, *Rand Journal of Economics* 19, 623–632.

Table I: Quality of Excess Valuation Calculations Across Methods (Conglomerates)

This table displays summary statistics for conglomerate benchmark valuations. Panel A is based on all conglomeratates, Panel B restricts attention to those with excess valuations within the interval $\{-1.386, +1.386\}$, and Panel C restricts attention to observations for which all methods generate excess valuations within this range (this holds the sample size fixed). The **Berger+Ofek Baseline** benchmarks are based on Berger and Ofek (1995). The **SIC Universe: Whole Firm, Unconstrained** benchmarks use text-based weights to construct the benchmarks. The **HP: SIC+TNIC Universe: Whole Firm, Unconstrained** benchmarks extend this method by expanding the set of available pure plays to include TNIC peers. The **HP: SIC+TNIC Universe (wf): Whole Firm, Constrained** benchmarks extend this method further using constrained regression to match the conglomerate on five accounting characteristics. The **HP: SIC+TNIC Universe: Constrained, Segment-by-Segment** benchmarks additionally account for segment-by-segment sales.

Row	Benchmark	Excess Value (Sales Based)	MSE Excess Val. (Sales based)	# Obs. (Sales based)	Excess Value (Assets Based)	MSE Excess Val. (Assets based)	# Obs. (Assets based)	Std. Dev. Weights
	Panel A: Raw Data							
1	Berger+Ofek Baseline (ss)	-0.082	0.474	12714	-0.027	0.288	10916	
2	HP: SIC Universe (wf): Unconstrained	-0.058	0.463	12714	-0.038	0.268	12714	0.041
3	HP: SIC+TNIC Universe (wf): Unconstrained	-0.046	0.402	12733	-0.008	0.242	12733	0.031
4	HP: SIC+TNIC Universe (wf): Constrained	-0.012	0.320	12773	-0.001	0.208	12773	0.047
5	HP: SIC+TNIC Universe (ss): Constrained, Segment-by-Segment	-0.018	0.377	12675	0.020	0.282	10902	0.058
	Panel B: Restrict to Excess Valuations to interval [-1.386,+1.386] (Berger and Ofek)							
6	Berger+Ofek Baseline (ss)	-0.069	0.334	11892	-0.066	0.212	8761	
7	HP: SIC Universe (wf): Unconstrained	-0.047	0.342	11912	-0.033	0.216	8805	0.041
8	HP: SIC+TNIC Universe (wf): Unconstrained	-0.038	0.314	12079	-0.014	0.194	8823	0.031
9	HP: SIC+TNIC Universe (wf): Constrained	-0.012	0.252	12213	-0.009	0.166	8844	0.047
10	HP: SIC+TNIC Universe (ss): Constrained, Segment-by-Segment	-0.012	0.281	12053	-0.017	0.191	8744	0.058
	Panel C: Uniformly Restrict to interval [-1.386,+1.386]							
11	Berger+Ofek Baseline (ss)	-0.065	0.306	11152	-0.049	0.183	7716	
12	HP: SIC Universe (wf): Unconstrained	-0.040	0.308	11152	-0.018	0.190	7748	0.041
13	HP: SIC+TNIC Universe (wf): Unconstrained	-0.028	0.274	11152	-0.001	0.171	7766	0.030
14	HP: SIC+TNIC Universe (wf): Constrained	0.004	0.210	11152	0.002	0.143	7778	0.045
15	HP: SIC+TNIC Universe (ss): Constrained, Segment-by-Segment	0.000	0.244	11152	-0.003	0.169	7720	0.056

Table II: Quality of Excess Valuation Calculations Across Methods (Single Segment Firms)

This table displays summary statistics for single segment benchmark valuations. Panel A is based on all single segment firms, Panel B restricts attention to those with excess valuations within the interval $\{-1.386, +1.386\}$, and Panel C restricts attention to observations for which all methods generate excess valuations within this range (this holds the sample size fixed). The **Berger+Ofek Baseline** benchmarks are based on Berger and Ofek (1995). The **SIC Universe: Whole Firm, Unconstrained** benchmarks use text-based weights to construct the benchmarks. The **HP: SIC+TNIC Universe: Whole Firm, Unconstrained** benchmarks extend this method by expanding the set of available pure plays to include TNIC peers. The **HP: SIC+TNIC Universe (wf): Whole Firm, Constrained** benchmarks extend this method further using constrained regression to match the conglomerate on five accounting characteristics.

Row	Benchmark	Excess Value (Sales Based)	MSE Excess Val. (Sales based)	# Obs. (Sales based)	Excess Value (Assets Based)	MSE Excess Val. (Assets based)	# Obs. (Assets based)	Std. Dev. Weights
	Panel A: Raw Data							
1	Berger+Ofek Baseline (ss)	-0.047	0.602	37579	0.045	0.339	37578	
2	HP: SIC Universe (wf): Unconstrained	0.012	0.579	37575	0.051	0.354	37574	0.053
3	HP: SIC+TNIC Universe (wf): Unconstrained	-0.019	0.507	37583	0.044	0.320	37582	0.028
4	HP: SIC+TNIC Universe (wf): Constrained	0.001	0.417	37783	0.041	0.281	37782	0.066
	Panel B: Restrict to Excess Valuations to interval [-1.386,+1.386] (Berger and Ofek)							
5	Berger+Ofek Baseline (ss)	-0.017	0.356	34686	0.022	0.263	36663	
6	HP: SIC Universe (wf): Unconstrained	0.009	0.343	34772	0.031	0.268	36419	0.053
7	HP: SIC+TNIC Universe (wf): Unconstrained	-0.006	0.331	35295	0.025	0.251	36690	0.028
8	HP: SIC+TNIC Universe (wf): Constrained	0.001	0.276	35948	0.026	0.217	36835	0.065
	Panel C: Uniformly Restrict to interval [-1.386,+1.386]							
9	Berger+Ofek Baseline (ss)	-0.026	0.316	32384	0.034	0.215	31955	
10	HP: SIC Universe (wf): Unconstrained	0.012	0.308	32384	0.040	0.228	31929	0.052
11	HP: SIC+TNIC Universe (wf): Unconstrained	-0.007	0.284	32384	0.033	0.212	31983	0.028
12	HP: SIC+TNIC Universe (wf): Constrained	0.009	0.232	32384	0.032	0.187	32076	0.058

Table III: Characteristic Correlations (Conglomerate Single Segment Firms vs. Replicating Peers)

The table displays Pearson Correlation coefficients between actual focal firm characteristics and implied characteristics of replicating peers. We consider several different replicating peer methods as noted in the column headers. Panel A reports results for conglomerate focal firms and Panel B reports results for single segment focal firms.

Row	Variable	Berger + Ofek (Baseline)	Text-based SIC only No Constr.	Text-based SIC+TNIC No Constr.	Text-based SIC+TNIC Constrained	Text-based SIC+TNIC Constrained (Seg by Seg)
			Panel A: Correlation Coefficients: Conglomerates			
1	Assets	0.110	0.194	0.291	0.409	0.399
2	Sales	0.156	0.229	0.385	0.387	0.315
3	OI/Sales	0.375	0.425	0.479	0.850	0.675
4	OI/Assets	0.289	0.357	0.421	0.832	0.690
5	R&D/Sales	0.473	0.673	0.705	0.908	0.821
6	Tobin's Q	0.366	0.442	0.469	0.551	0.502
7	Sales Growth	0.241	0.270	0.309	0.825	0.683
8	Log Age	0.268	0.298	0.436	0.924	0.731
			Panel B: Correlation Coefficients: Single Segment Firms			
9	Assets	0.100	0.063	0.282	0.363	N/A
10	Sales	0.204	0.184	0.368	0.349	N/A
11	OI/Sales	0.402	0.463	0.508	0.833	N/A
12	OI/Assets	0.131	0.190	0.203	0.499	N/A
13	R&D/Sales	0.403	0.594	0.637	0.845	N/A
14	Tobin's Q	0.225	0.330	0.295	0.500	N/A
15	Sales Growth	0.316	0.308	0.360	0.825	N/A
16	Log Age	0.330	0.283	0.392	0.888	N/A

Table IV: Which Replicating Peers Match with Conglomerates and Single Segment Firms?

The table displays summary statistics for replicating peers assigned above median weights versus below median weights for conglomerate focal firms (Panel A), and single segment focal firms (Panel B).

		Benchmark Portfolio Weights vs Characteristics					
		SIC+TNIC Universe: Whole Firm, Un-constrained			*SIC+TNIC Universe: Whole Firm, Constrained*		
Row	Variable	Lowest Weights Quartile	Highest Weights Quartile	*t*-statistic of Difference	Lowest Weights Quartile	Highest Weights Quartile	*t*-statistic of Difference
				Panel A: Conglomerates			
1	Assets	3466.56	4723.34	6.52	3564.72	4934.27	6.85
2	Sales	1563.12	2147.15	9.49	1580.38	2213.75	10.53
3	oi/sales	0.07	0.08	6.59	0.07	0.08	8.00
4	oi/assets	0.07	0.07	0.55	0.07	0.07	1.20
5	R+D/sales	0.11	0.09	-14.15	0.11	0.09	-14.20
6	Tobin's Q	2.05	1.92	-4.84	2.04	1.88	-6.50
7	Sales Growth	0.17	0.16	-9.10	0.18	0.16	-17.92
8	Firm Age	25.32	29.19	18.63	24.06	30.32	29.14
				Panel B: Single Segment Firms			
9	Assets	7305.50	6584.34	-1.51	7481.07	7023.68	-3.13
10	Sales	1437.45	1392.45	-0.76	1473.84	1401.71	-2.46
11	OI/Sales	0.15	0.15	0.35	0.15	0.15	5.72
12	OI/ssets	0.05	0.05	0.78	0.05	0.05	1.51
13	R&D/Sales	0.08	0.08	-2.38	0.08	0.07	-5.53
14	Tobin's Q	1.44	1.42	-2.05	1.43	1.40	-2.70
15	Sales Growth	0.15	0.15	-4.28	0.15	0.15	-4.77
16	Firm Age	23.90	24.29	2.31	24.28	23.94	-3.78

Table V: Conglomerate and Single Segment Firm Excess Valuations

OLS regressions with time fixed effects and standard errors clustered by firm. One observation is one conglomerate focal firm from 1997 to 2008. The dependent variable is the conglomerate focal firm's excess valuation computed using the best text-based reconstruction (Panel A) or using the Berger and Ofek reconstruction (Panel B). The best text-based reconstruction is the "HP: SIC+TNIC Universe: Constrained" model as illustrated in Table I. All independent variables are standardized to have a standard deviation of one for ease of interpretation and comparison of coefficients.

Row	Difficulty of Pure Plays to Replicate	Log Document Length	R&D/ Sales	CAPX/ Sales	OI/ Sales	Log Assets	# Obs. / RSQ
			Panel A: Conglomerate Excess Values (Text-based Constrained Valuation Model)				
(1)	0.066	-0.010	0.057	0.054	0.090	0.085	11,390
	(8.26)	(-1.21)	(6.76)	(7.08)	(9.45)	(8.92)	0.094
			Panel B: Conglomerate Excess Values (Berger + Ofek Valuation Model)				
(2)	0.085	0.027	0.107	0.066	0.134	0.131	11,077
	(8.34)	(2.64)	(10.90)	(6.35)	(11.60)	(11.46)	0.159
			Panel C: Single Segment Excess Values (Text-based Constrained Valuation Model)				
(3)	0.046	0.003	0.063	0.055	0.055	0.088	32,825
	(8.55)	(0.62)	(10.37)	(12.96)	(8.83)	(15.88)	0.055
			Panel B: Single Segment Excess Values (Berger + Ofek Valuation Model)				
(4)	0.067	0.037	0.212	0.108	0.167	0.140	31,440
	(11.08)	(6.31)	(26.64)	(20.62)	(19.22)	(20.85)	0.190

Table VI: Economic Magnitudes and Excess Valuation

This table displays average excess valuations for quintiles based on the difficulty of replicating peers to replicate. Panel A displays statistics for conglomerate focal firms, and Panel B displays results for single segment focal firms. For each quintile, we report the average difficulty variable, and average raw excess valuations based on both the"HP: SIC+TNIC Universe: Constrained" and Berger and Ofek methods. Residual excess valuations are residuals from a regression of excess valuation on all of the variables included in Table V and Table **??** excluding the Difficulty to Replicate variable.

Difficulty to Replicate Quintile	Difficulty to Replicate	Raw Excess Valuation (text-based)	Raw Excess Valuation (Berger+Ofek)	Residual Excess Valuation (text-based)	Residual Excess Valuation (Berger+Ofek)	Obs.
			Summary Statistics by Quintile			
			Panel A: Conglomerates			
Lowest Difficulty	0.630	-0.045	-0.035	-0.059	-0.058	2,331
Quintile 2	0.729	-0.003	-0.043	0.002	-0.009	2,339
Quintile 3	0.795	-0.007	-0.093	-0.000	-0.027	2,337
Quintile 4	0.858	-0.004	-0.094	0.015	-0.008	2,339
Highest Difficulty	1.028	0.047	-0.029	0.085	0.107	2,333
			Panel B: Single Segment Firms			
Lowest Difficulty	0.454	-0.007	0.027	-0.060	-0.065	6,782
Quintile 2	0.643	0.019	0.059	-0.009	-0.000	6,789
Quintile 3	0.725	-0.000	-0.054	0.000	-0.041	6,791
Quintile 4	0.809	0.016	-0.078	0.038	0.010	6,789
Highest Difficulty	1.103	0.032	-0.060	0.079	0.096	6,785

Table VII: Return Comovement

Fama-MacBeth regressions with own-firm monthly stock return as the dependent variable. One observation is one firm from 1997 to 2008. The independent variable includes the SIC-based return benchmark (excluding the firm itself) and the text-based return benchmark (also excluding the firm itself). We also include controls for the Fama and French (1992) variables (log book to market ratio and log size), a dummy for negative book to market ratio stocks (the dummy is not displayed to conserve space and is not significant), and a control for momentum (defined as the own-firm 11 month lagged return from month $t-12$ to $t-2$). Panel A displays results for all firms, and Panels B and C display results for conglomerates and single segment firms, respectively. Panel D displays results for all firms, however, we consider various lags to the dependent variable as noted in the dependent variable column. All peer variables are standardized to have a standard deviation of one for ease of comparison and interpretation.

Row	Dependent Variable	Text-based Peer Return	SIC-3 Peer Return	Log B/M Ratio	Log Size	Past 11 Mon. Return	# Obs. / RSQ
			Panel A: All Firms				
(1)	Month t Returns	0.046	0.014	0.002	-0.001	-0.002	582,907
		(45.33)	(22.93)	(2.33)	(-1.81)	(-0.64)	0.079
(2)	Month t Returns	0.055	.	0.002	-0.001	-0.002	582,907
		(54.21)		(2.24)	(-1.81)	(-0.63)	0.076
(3)	Month t Returns	.	0.036	0.002	-0.001	-0.002	582,907
			(38.62)	(1.67)	(-1.67)	(-0.41)	0.057
			Panel B: Conglomerate Firms Only				
(4)	Month t Returns	0.043	0.012	0.002	-0.001	-0.002	174,352
		(34.13)	(16.94)	(2.27)	(-1.64)	(-0.55)	0.079
(5)	Month t Returns	0.049	.	0.002	-0.001	-0.002	174,352
		(38.61)		(2.23)	(-1.63)	(-0.54)	0.075
(6)	Month t Returns	.	0.030	0.002	-0.001	-0.001	174,352
			(28.49)	(1.94)	(-1.55)	(-0.33)	0.056
			Panel C: Single Segment Firms Only				
(7)	Month t Returns	0.047	0.014	0.002	-0.001	-0.002	408,555
		(42.66)	(18.95)	(2.27)	(-1.72)	(-0.68)	0.082
(8)	Month t Returns	0.056	.	0.002	-0.001	-0.002	408,555
		(53.76)		(2.18)	(-1.72)	(-0.66)	0.079
(9)	Month t Returns	.	0.038	0.002	-0.001	-0.002	408,555
			(37.73)	(1.56)	(-1.60)	(-0.47)	0.061
			Panel D: All Firms, various lags				
(10)	Month t Returns	0.046	0.014	0.002	-0.001	-0.002	582,907
		(45.33)	(22.93)	(2.33)	(-1.81)	(-0.64)	0.079
(11)	Month $t+1$ Returns	0.012	0.003	0.001	-0.001	-0.003	578,025
		(6.19)	(4.05)	(1.23)	(-1.42)	(-0.83)	0.039
(12)	Month $t+2$ Returns	0.005	0.001	0.002	-0.001	-0.005	573,223
		(2.25)	(1.51)	(1.44)	(-1.34)	(-1.32)	0.038
(13)	Month $t+3$ Returns	0.003	0.001	0.001	-0.001	-0.005	568,478
		(1.66)	(1.50)	(0.75)	(-1.22)	(-1.30)	0.036

Table VIII: Return Comovement (High and Low Difficulty to Replicate)

Fama-MacBeth regressions with own-firm monthly stock return as the dependent variable. One observation is one firm from 1997 to 2008. The independent variable includes the SIC-based return benchmark (excluding the firm itself) and the text-based return benchmark (also excluding the firm itself). We also include controls for the Fama and French (1992) variables (log book to market ratio and log size), a dummy for negative book to market ratio stocks (the dummy is not displayed to conserve space and is not significant), and a control for momentum (defined as the own-firm 11 month lagged return from month $t-12$ to $t-2$). Panel A displays results for all firms, and Panels B and C display results for firms with high difficulty to replicate and low difficulty to replicate, respectively. All peer variables are standardized to have a standard deviation of one for ease of comparison and interpretation.

Row	Dependent Variable	Text-based Peer Return	SIC-3 Peer Return	Log B/M Ratio	Log Size	Past 11 Mon. Return	# Obs. / RSQ
			Panel A: All Firms				
(1)	Month t Returns	0.046	0.014	0.002	-0.001	-0.002	582,907
		(45.33)	(22.93)	(2.33)	(-1.81)	(-0.64)	0.079
(2)	Month $t+1$ Returns	0.012	0.003	0.001	-0.001	-0.003	578,025
		(6.19)	(4.05)	(1.23)	(-1.42)	(-0.83)	0.039
(3)	Month $t+2$ Returns	0.005	0.001	0.002	-0.001	-0.005	573,223
		(2.25)	(1.51)	(1.44)	(-1.34)	(-1.32)	0.038
(4)	Month $t+3$ Returns	0.003	0.001	0.001	-0.001	-0.005	568,478
		(1.66)	(1.50)	(0.75)	(-1.22)	(-1.30)	0.036
			Panel B: Above Median Difficulty to Replicate Only				
(5)	Month t Returns	0.032	0.009	0.002	-0.001	-0.003	271,497
		(30.67)	(13.33)	(2.69)	(-1.52)	(-0.84)	0.052
(6)	Month $t+1$ Returns	0.008	0.002	0.002	-0.001	-0.005	269,281
		(6.39)	(3.42)	(2.20)	(-1.28)	(-1.25)	0.030
(7)	Month $t+2$ Returns	0.003	0.001	0.002	-0.001	-0.007	267,109
		(2.17)	(1.01)	(2.24)	(-1.19)	(-1.80)	0.029
(8)	Month $t+3$ Returns	0.003	0.001	0.002	-0.001	-0.006	264,968
		(2.06)	(1.62)	(1.80)	(-1.12)	(-1.75)	0.028
			Panel C: Below Median Difficulty to Replicate Only				
(9)	Month t Returns	0.065	0.018	0.001	-0.001	-0.002	271,976
		(48.93)	(16.24)	(1.37)	(-1.91)	(-0.48)	0.123
(10)	Month $t+1$ Returns	0.018	0.002	0.000	-0.001	-0.001	269,603
		(5.76)	(1.82)	(0.26)	(-1.23)	(-0.38)	0.061
(11)	Month $t+2$ Returns	0.006	0.001	0.001	-0.001	-0.003	267,274
		(1.89)	(1.00)	(0.51)	(-1.32)	(-0.94)	0.060
(12)	Month $t+3$ Returns	0.003	0.001	-0.000	-0.001	-0.003	264,977
		(1.01)	(0.85)	(-0.22)	(-1.21)	(-0.84)	0.057

Table IX: Return Comovement (Systematic versus Idiosyncratic Components)

Fama-MacBeth regressions with own-firm monthly stock return as the dependent variable. One observation is one firm from 1997 to 2008. The independent variables include the systematic and idiosyncratic portions of the text-based return benchmark. To compute the systematic portion, we first regress (for each month) daily stock returns for each firm onto the three Fama French factors and the momentum factor. The projection from this regression (excluding the projection from the intercept) is the systematic portion of a firm's daily return. These are then aggregated to monthly observations, and we compute the average of these systematic returns over each firm's text based peers to get the "Systematic Peer Return". The idiosyncratic Peer Return is the raw text-based peer return minus the systematic peer return. Panel A displays results for all firms, and Panels B to E display results for various subsamples as noted. Peer variables are standardized to have a standard deviation of one for ease of comparison and interpretation.

Row	Dependent Variable	Systematic Peer Return	Idio. Peer Return	Log B/M Ratio	Log Size	Past 11 Mon. Return	# Obs. / RSQ
			Panel A: All Firms				
(1)	Month t Returns	0.053	0.033	0.002	-0.001	-0.002	582,907
		(38.34)	(51.80)	(2.21)	(-1.78)	(-0.64)	0.078
(2)	Month $t+1$ Returns	0.014	0.009	0.001	-0.001	-0.003	578,025
		(4.16)	(7.36)	(1.23)	(-1.50)	(-0.84)	0.043
(3)	Month $t+2$ Returns	0.003	0.003	0.002	-0.001	-0.005	573,223
		(0.70)	(2.77)	(1.56)	(-1.39)	(-1.29)	0.042
(4)	Month $t+3$ Returns	-0.000	0.003	0.001	-0.001	-0.005	568,478
		(-0.00)	(2.68)	(1.02)	(-1.13)	(-1.41)	0.041
			Panel B: Above Median Difficulty to Replicate Only				
(5)	Month t Returns	0.039	0.023	0.002	-0.001	-0.003	271,497
		(25.27)	(35.03)	(2.66)	(-1.52)	(-0.85)	0.052
(6)	Month $t+1$ Returns	0.009	0.006	0.002	-0.001	-0.005	269,281
		(3.46)	(7.58)	(2.18)	(-1.32)	(-1.27)	0.032
(7)	Month $t+2$ Returns	0.001	0.002	0.002	-0.001	-0.007	267,109
		(0.43)	(2.66)	(2.31)	(-1.24)	(-1.82)	0.031
(8)	Month $t+3$ Returns	0.001	0.003	0.002	-0.001	-0.007	264,968
		(0.36)	(3.06)	(2.04)	(-1.03)	(-1.82)	0.031
			Panel C: Below Median Difficulty to Replicate Only				
(9)	Month t Returns	0.069	0.049	0.001	-0.001	-0.002	271,976
		(43.70)	(62.85)	(1.21)	(-1.90)	(-0.51)	0.120
(10)	Month $t+1$ Returns	0.021	0.012	0.000	-0.001	-0.001	269,603
		(4.17)	(6.39)	(0.32)	(-1.40)	(-0.40)	0.068
(11)	Month $t+2$ Returns	0.004	0.005	0.001	-0.001	-0.003	267,274
		(0.62)	(2.38)	(0.54)	(-1.31)	(-0.85)	0.066
(12)	Month $t+3$ Returns	-0.003	0.004	-0.000	-0.001	-0.004	264,977
		(-0.53)	(2.06)	(-0.07)	(-1.13)	(-1.00)	0.065
			Panel D: Conglomerates Only				
(13)	Month t Returns	0.050	0.030	0.002	-0.001	-0.002	174,352
		(27.44)	(37.03)	(2.21)	(-1.62)	(-0.58)	0.077
(14)	Month $t+1$ Returns	0.010	0.007	0.002	-0.001	-0.003	173,185
		(3.13)	(6.21)	(1.90)	(-1.45)	(-0.72)	0.048
(15)	Month $t+2$ Returns	0.002	0.004	0.003	-0.001	-0.004	172,026
		(0.60)	(2.88)	(2.14)	(-1.23)	(-1.14)	0.046
(16)	Month $t+3$ Returns	0.001	0.003	0.002	-0.001	-0.005	170,868
		(0.18)	(2.52)	(1.70)	(-1.08)	(-1.27)	0.045
			Panel E: Single Segment Firms Only				
(17)	Month t Returns	0.054	0.034	0.001	-0.001	-0.002	408,555
		(35.32)	(50.34)	(2.09)	(-1.68)	(-0.66)	0.081
(18)	Month $t+1$ Returns	0.016	0.009	0.001	-0.001	-0.003	404,840
		(4.41)	(7.20)	(0.99)	(-1.42)	(-0.89)	0.046
(19)	Month $t+2$ Returns	0.003	0.003	0.001	-0.001	-0.005	401,197
		(0.64)	(2.58)	(1.34)	(-1.35)	(-1.39)	0.044
(20)	Month $t+3$ Returns	-0.000	0.003	0.001	-0.001	-0.005	397,610
		(-0.09)	(2.49)	(0.80)	(-1.05)	(-1.49)	0.043