Conglomerate Industry Choice and Product Differentiation

Gerard Hoberg and Gordon Phillips[*]

February 25, 2012

ABSTRACT

We use text-based computational analysis of business descriptions from 10-Ks to examine in which industries conglomerates are most likely to operate and to understand conglomerate valuations. We find that conglomerates are more likely to operate in industry pairs that are closer together in the product space and in industry pairs that have profitable opportunities "between" them. Conglomerate firms have lower stock market valuations than matched single-segment firms when their products are easier to replicate with single-segment firms. Conglomerate firms have stock market premiums when they have higher product differentiation and produce in more profitable industries. These findings are consistent with successful conglomerate firms having higher product differentiation and lower cost entry into profitable markets when operating in strategically chosen industry pairs.

[*]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. This paper entirely replaces a previous version titled "Conglomerate Industry Spanning". We thank Merih Sevilir, Denis Sosyura and seminar participants at Rotterdam School of Management, Erasmus University, Stanford University, Tilberg University, University of Illinois and the University of Mannheim for helpful comments. All errors are the authors alone.

Why do firms operate in multiple industries? This question has been the focus of much research that takes the industries that firms operate in as given. Existing explanations for multiple industry production focus on examining ex post investment and productivity in an attempt to understand differences between conglomerate firms and single-segment firm valuations.[1] However, this literature does not examine why conglomerate firms choose some industry combinations and not others and does not explain the cross-sectional variation in conglomerate valuations based on industry and product choices.

We take a different approach in this study. We use fundamental product market characteristics to analyze in which industry combinations conglomerate firms operate and how these industry and product choices relate to valuations. Our central hypotheses are related to theories of economies of scope (Hay (1976) and Panzar and Willig (1981)), and product differentiation (Chamberlin (1933), Berry, Levinsohn, and Pakes (1997), and Seim (2006)). We expand on these ideas in two novel ways, both relating to conglomerate structure. First, we explore whether conglomerates choose to operate in industry pairs that have profitable industries "between" them, consistent with reducing potential entry costs into related markets. Second, we examine whether conglomerates are more valuable when they cannot be easily reconstructed using single-segment firm portfolios, consistent with successful multiple-industry firms adding unique differentiated products related to their industry choices that are not produced by single-segment rivals.

Apple Computer is an example of a firm which illustrates our ideas. In this case its multiple industry structure likely enables it to produce differentiated products competing with cell phones, computers, and digital music - industries that are very related today. Apple was successful in its decision to operate jointly in these industries. Given that much literature focuses on the potentially harmful aspects of the conglomerate multiple-industry structure, our paper thus sheds new light on the

[1] Stein (1997))and Khanna and Tice (2001) focuses on the advantages of internal capital markets in allocating investment capital, Shin and Stulz (1998), Denis, Denis, and Sarin (1997), and Scharfstein and Stein (2000) focus on the effect of agency problems in allocating investment, Matsusaka (2001) and Hund, Monk, and Tice (2010) focus on conglomerate learning about ability, Maksimovic and Phillips (2002) focus on effect of differences in managerial talent on investment and productivity and Schoar (2002) focuses on the effect of managerial attention on productivity.

puzzling question: Why do so many firms continue to use the conglomerate structure?

We use text-based analysis of conglomerate and pure play business descriptions from 10-Ks filed with the SEC to examine in which industries conglomerates are most likely to operate and to explore conglomerate valuations. Following Hoberg and Phillips (2010a), we convert firm product text into a spatial representation of the product market. In this framework, each firm and each industry has a product location based on words that allow measurement of how close product markets are to each other. Our spatial framework also allows an assessment of how similar industries are to each other (across-industry similarity), how similar firms in an industry are to each other (within-industry similarity), and which industries in the product market space are "between" any given pair of industries.[2] We control for other measures of relatedness including vertical integration in assessing the impact of our text-based measures of relatedness.

We find that conglomerates are more likely to operate in industry pairs that have high across-industry similarity, lower within-industry similarity, and when industry pairs have profitable, less contested opportunities between them. These findings are consistent with product market synergies and economies of scope, and suggest that conglomerates producing in two related industries may be able to enter another profitable industry that lies between them at low cost. The focus on industries with low within-industry similarity suggests that the potential for product differentiation also likely plays a role in industry choice.

We next consider how industry choice and product differentiation relate to conglomerate valuations using firm product text to find best matches for each conglomerate firm. We use least squares vocabulary reconstructions to find weights for the best matching single-segment firms such that the weighted sum of their individual product vocabularies best matches that of the given conglomerate. Single-segment firms receiving higher weights offer products that are more central to explaining a given conglomerate's observed product offerings.[3] We show that our weighted benchmarks

[2]"Between" industries are industries that are closer to each industry of a given industry pair than the industry pair is to each other based on product text similarity. We formally define this measure in the next section.

[3]These "network" benchmarks represent best matches in the product market analogous to a

provide economically large improvements relative to existing methods in their ability to accurately match conglomerate valuations and characteristics. We use these new benchmarks and measures of product differentiation to examine cross-sectional differences in conglomerate valuations - not to test for an average conglomerate discount which disappears using our text-matched single-segment firms.[4]

Our main valuation results show that the median or average valuation of conglomerate firms masks important cross-sectional variation. Conglomerates with higher product differentiation, measured by the difficulty to reconstruct using single segment firms, trade at stock-market premia. Those with less differentiation that are better matched or "spanned" by single-segment firms trade at discounts. We also find that conglomerate valuations are higher when conglomerates have high-value industries between their operating segment industries. These findings are consistent with successful conglomerate firms creating differentiated products that are more difficult to replicate by single segment firms, and with a successful conglomerate organization lowering the cost of entry into valuable industries.

Our paper makes three main contributions. First, our paper examines in which industry combinations conglomerate firms choose to operate. We find that across-industry similarity, within-industry similarity, and the nature of industries lying between two industries explain conglomerate industry choice. Our second contribution is methodological: we present new text-based methods for generating single-segment benchmarks for conglomerate valuation. These benchmarks offer significant gains in accuracy relative to existing methods. Our third contribution is to examine the link between product market variables and conglomerate valuations in cross section. We find that conglomerate valuations are higher when the conglomerate is more difficult to replicate using single-segment firms, and when its segments have high-value industries between them. Given our focus is on understanding the cross sectional valuation

Facebook circle of friends (both close friends and acquaintances).

[4]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), Rajan, Servaes, and Zingales (2000) and Lamont and Polk (2002) find evidence of a diversification discount. However 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.

of conglomerates, it fills a gap empirically that Stein (2003) identifies in his survey paper.[5] In all, our findings support theoretical links to economies of scope, product differentiation, and product market synergies. Our results also help to explain why so many firms continue to use the conglomerate structure despite potential negative effects on valuation noted by past studies.

Our paper proceeds as follows. In the next section, we present new measures of industry relatedness and we develop our key hypotheses. In Section II, we discuss our data, variables, and methods used to examine industry choice. Section III presents the results of our analysis of industry choice. Section IV presents our methodology for constructing conglomerate benchmarks using pure play firms, and Section V analyzes cross-sectional conglomerate valuations. Section VI concludes.

I Industry Similarity and Relatedness

We ask whether there are certain industry characteristics - distinct from vertical relatedness - that make operating in two different industries valuable. The central hypothesis we examine is whether the potential for product differentiation, product market synergies, and low cost entry into "between industries" influence in which industries conglomerate firms operate. Our foundation draws upon literature relating to economies of scopeHay (1976) and Panzar and Willig (1981)), and product differentiation (Chamberlin (1933), Berry, Levinsohn, and Pakes (1997), and Seim (2006)) and product market synergies (Rhodes-Kropf and Robinson (2008)).

We generate industry pair characteristics using text-based analysis of business descriptions from 10-Ks filed with the SEC. We then examine these industry characteristics to understand in which industries conglomerates are most likely to operate and to understand cross-sectional conglomerate valuation. We discuss the way we gather and process these 10-K product descriptions in the next section. In this section we introduce conceptually the variables we use to capture how industries are related to each other. We consider these new measures in addition to existing

[5]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 and empirically.

industry-relatedness measures including vertical integration.

To construct industry relatedness measures, we begin by using the relatedness of each pairwise set of firms that operate either within an industry or across any given industry pair. These product relatedness measures are constructed for each pair of firms using the words from each firm's business description from their 10K filed with the SEC. We discuss the specific word relatedness measures we use - the cosine similarity measure - later in the subsequent data and methodology section.

We construct three new measures of industry relatedness that allow us to assess how every pair of industries relates to one another. In particular, we measure how far apart industries are in the product space, *Across Industry Similarity (AIS)*, how heterogeneous their products are within-industry, *Within Industry Similarity (WIS)*, and the extent to which other industries lie between the given industry pair in the product space, *Between Industries (BI)*. These measures are as follows:

Across Industry Similarity (AIS): This measure captures the extent that product descriptions of firms in two different industries are similar. The AIS measure is meant to capture the similarities between products that two industries produce. We examine this measure as industries that are closer together, in the sense of sharing related products, are likely to share operating synergies and asset complementarities. Specifically, across industry similarity is the average textual cosine similarity of all pairwise permutations of the N_i and N_j firms, where textual similarity is based on word vectors from firm business descriptions (see Section II.C for a discussion of the cosine similarity method). Simply put, it captures the proportion of product words the two firms in a pair have in common.

Within Industry Similarity (WIS): This measure captures the product differentiation within an industry as within industry similarity is the average cosine similarity of the business descriptions for all pairwise permutations of these N_i firms. We consider whether within industry similarity decreases the incentives for firms to operate in a particular industry as firms in industries with higher within-industry similarity are likely to have less unique products, and likely face more significant competition from their rivals due to the absence of product differentiation. If there are additional costs in setting up and operating firms with a multiple industry structure, the costs may likely to outweigh the benefits of operating in industries with high within industry similarity.

Between Industries (BI): We use the across industry measures to assess what industries are between any given industry pair. Specifically, an industry is between two industries in a given industry pair if it is closer in textual distance to each industry in the pair than they are to each other. We hypothesize that a conglomerate is more likely to produce in a particular industry pair if that pair has other highly valued, less competitive industries, between the pair. Producing in such an industry pair may allow multiple industry firms to more easily enter and produce products in these highly-valued concentrated product markets. The AIS measure discussed above is instrumental in computing the fraction of industries between a given pair. Where $AIS_{i,j}$ denotes the Across Industry Similarity of industries i and j, we define a third industry k as being *between* industries i and j if the following relationship holds.

$$AIS_{k,i} \leq AIS_{i,j} \qquad \text{AND} \qquad AIS_{k,j} \leq AIS_{i,j} \tag{1}$$

The fraction of industries between a given pair of industries i and j is therefore the number of industries k (excluding i and j) satisfying this condition divided by the total number of industries in the database in the given year (excluding i and j).

Given we will be examining in which industries conglomerate firms produce, we focus on single-segment firms to calculate these industry relatedness measures. We exclude conglomerates and their product descriptions from the calculation of relatedness measures to avoid producing a mechanistic relation. We then use the conglomerate firm's Compustat segment definitions to examine how the observed conglomerate industry configurations and the words conglomerate firms use in their product description relate to these text-based industry attributes from single-segment firms.

We use these new industry relatedness measures from firm product text to test the following three hypotheses:

H1: Cross-Industry Similarity: Conglomerate firms are more likely to produce in two industries that have high cross-industry similarity. These firms are likely easier to manage and have more potential for product market synergies.

We test this hypothesis by examining the number of conglomerate firms that operate in each pairwise set of industries and examine whether this number of conglomerate firms is increasing in the pair's across-industry similarity (AIS).

H2: Within-Industry Similarity: Conglomerate firms are less likely to produce in industries that have high within-industry similarity, or industries with little potential for product differentiation.

We test this hypothesis by examining whether the number of conglomerate firms operating in a pair is decreasing when the industry pair has a high within-industry similarity (WIS).

H3: Between-Industries: Conglomerate firms are more likely to operate in an industry pair when the pair of industries has more high-value, less competitive industries, between the pair.

We examine the fraction of industries that are between each pairwise combination of industries and test whether conglomerate firms producing in a particular pairwise combination increase when the industries between these industries are highly valued and less competitive.

H4: Conglomerate Valuations and Spanning: Conglomerate valuation will be higher when the conglomerate firm is not easily matched or "spanned" by single-industry segment firms (i.e., the conglomerate is difficult to replicate and has products that exhibit product differentiation relative to single-segment firms). Conglomerate valuation will also be higher when the firm has more highly valued industries between its industry pairs.

II Data and Methodology

In this section we describe our conglomerate database, the construction of key text-based variables used to examine where conglomerates produce in the product space, and our identification of single-segment (also called pure-play) conglomerate competitors.

A The COMPUSTAT Industry Sample

We construct our COMPUSTAT sample using the industrial annual files to identify the universe of publicly traded firms, and the COMPUSTAT segment files to identify which firms are conglomerates, and the industry of each segment. We define a conglomerate as a firm having operations in more than one SIC-3 industry in a given year. To identify segments operating under a conglomerate structure, we start with the segment files, which we clean to ensure we are identifying product-based segments instead of geographic segments. We keep conglomerate segments that are identified as business segments or operating segments. We only keep segments which report positive sales. We aggregate segment information into 3 digit SIC codes and only identify firms as conglomerate firms when they report two or more three digit SIC codes. We identify 22,252 unique conglomerate firm years from 1996 to 2008 (we limit our sample to these years due to required coverage of text-based variables), which have 62,058 unique conglomerate-segment-years. We also identify 56,491 unique pure play firm-years (firms with a single segment structure).

When we examine how conglomerates change from year to year, we further require that a conglomerate exist in the previous year. This requirement reduces our sample to 18,589 unique conglomerate years having 53,126 segment-years. Because we use pure play firms to assess industry characteristics that might be relevant to the formation of conglomerates, we also discard conglomerate observations if they have at least one segment operating in an industry for which there are no pure play benchmarks in our sample. We are left with 15,373 unique conglomerate firm-years with 40,769 unique segment conglomerate firm-years. This final sample covers 2,552 unique three digit SIC industry-years. As there are 13 years in our sample, this is roughly 196 industries per year.

We also consider a separate database of pairwise permutations of the SIC-3 industries in each year. We use this database to assess which industry pairs are most likely to be populated by conglomerates that operate in the given pair of industries. This industry-pair-year database has 312,240 total industry pair x year observations (roughly 24,018 industry pair permutations per year).

B 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 for text information and a company identifier, CIK.[6] 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.

C Word Vectors and Cosine Similarity

After we have 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.[7] 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, 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

[6]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.

[7]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 results were qualitatively similar.

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.

D Conglomerate Competitors

We use our text-based analysis of firms to redefine and augment traditional identification of conglomerate competitors, which allows us to construct new measures of excess conglomerate valuation. As a baseline calculation, we begin by following the existing literature (Lang and Stulz (1994) and Berger and Ofek (1995)) and we consider existing pure-play single-segment firms as identified the Compustat segment tapes as the set of competitors for each segment of a conglomerate. We refer to this set of competitors as the Compustat set of rivals, and this set includes single-segment firms with the same 3-digit SIC code as the reported Compustat segment. We then consider several enhancements. First, rather than computing valuation ratios as the simple median of all candidate single-segment rivals for a given conglomerate segment, we weight each single segment firm using information from a vocabulary decomposition of the conglomerate relative to the vocabulary used by all pure play candidate rivals in their business descriptions. We explain this weighting methodology in Section IV. Firms with more words in common with the conglomerate will have higher weights.

Our second augmentation is to expand the set of "pure-play" single-segment rivals by adding the single-segment firms that are in the conglomerate's "Text-Based Network Industry Classification" (TNIC) industry as defined in Hoberg and Phillips

(2010a). These firms have a high product word cosine similarity score relative to the conglomerate firm and are likely rivals. Importantly, the TNIC industry classification is equally as coarse as are SIC-3 industries, so our results are not due to any changes in industry coarseness. The number of potential competitor benchmarks for each conglomerate segment $N_{it,bench}$ is thus either as large (if no pure play TNIC peers exist) or larger (if pure play TNIC peers do exist) than the Compustat set of rivals. We refer to this method as the "SIC+TNIC Universe" universe.

After identifying the potential competitors for each segment, we again weight the potential firms based on the vocabulary decomposition to construct conglomerate valuation benchmarks and excess valuation measures. We also consider further enhancing the weighting scheme of the set of potential competitors so that pure play benchmarks can match the conglomerate on accounting characteristics including profitability and sales growth. We discuss these valuation metrics and how we use them to construct excess valuation measures for each conglomerate firm fully in Section IV when we examine the cross-sectional and time-series dimensions of conglomerate valuation.

E Conglomerate Restructuring

We examine whether our spatial industry variables can explain how conglomerates restructure over time, and we classify restructuring in three different ways. Because we consider the role of industry topography, the unit of observation for these variables is a pair of segments operating within a conglomerate. We define "Segment Pair Disappears" as a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year. We then define "Segment Pair Likely Sold or Closed" as a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate has fewer segments in year $t+1$ relative to year t. Finally, we define "Segment Pair Likely Sold Off" as a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate was the target of an acquisition of at least ten percent of its assets in year $t+1$. We define an analogous set of variables to assess conglomerate segment pairs that newly appear in year t that were not part of the conglomerate in year $t-1$.

F Control Variables and Vertical Integration

In addition to our three new industry similarity and relatedness variables, we include control variables for industry size, vertical relatedness, and a dummy identifying which industries are in the same two-digit SIC code. As we aim to examine conglomerate incidence rates across industry pairs, controlling for industry size is important. For example, if conglomerates formed by randomly choosing among available pure play firms in the economy, then the incidence of conglomerate operating pairs would be related to the product of the fraction of firms residing in industries i and j. Therefore we define the Pair Likelihood if Random variable as the product $(F_i x F_j)$, where F_i is the number of pure play firms in industry i divided by the number of pure play firms in the economy in the given year.

We consider the Input/Output tables to assess whether conglomerates tend to operate in vertically related industry pairs. The inclusion of this control is motivated by studies examining vertically related industries and corporate policy and structure including Fan and Goyal (2006), Kedia, Ravid, and Pons (2008), and Ahern and Harford (2011). We consider the methodology described in Fan and Goyal (2006) to identify vertically related industries. Based on three-digit SIC industries, we use the "Use Table" of Benchmark Input-Output Accounts of the US Economy to compute, for each firm pairing, the fraction of inputs that flow between each pair.

Finally, we consider a dummy variable set equal to one if a given pair of three digit SIC industries lies in the same two-digit SIC industry.

G Summary Statistics

Table I displays summary statistics for our conglomerate and pure play firms, and industry pair databases. Panel A shows that conglomerate firms are generally larger than the pure play firms in terms of total value of the firm, and they also generally operate in markets that are more concentrated, as measured by their TNIC HHI.

Panel B of the table compares randomly drawn pairs of SIC-3 industries to the

SIC-3 industries comprising a conglomerate configuration. The panel shows that a randomly drawn pair of three digit SIC industries has 0.147 conglomerates having segments operating in both industries of the given pair. Hence, the majority of randomly chosen industry pairs do not have conglomerates operating in the pair. The average across-industry similarity of random pairs is 0.017, which closely matches the average firm similarity reported in Hoberg and Phillips (2010a). This quantity is nearly double for actual conglomerates at 0.032, indicating that conglomerates are far less diversified than previously thought. This conclusion is reinforced by comparing the fraction of all other industries lying between the given pair, which is 32.5% for random pairs, and just 9.7% for actual conglomerates. Conglomerate industry pairs are in regions of the product space that are substantially closer together than randomly chosen industries. The average within-industry similarity, intuitively, is much higher at 0.086. This quantity is somewhat lower at 0.073 for actual conglomerates.

[**Insert Table I Here**]

Table II displays the bivariate Pearson correlation coefficients for our key industry pair variables. The key variable we examine in the next section is the number of conglomerates operating in a given pair. The first column of this table shows that this variable is positively related to across-industry similarity, and negatively related to within-industry similarity and the fraction of industries between a given pair. Although these univariate results hold for across-industry similarity and within-industry similarity, multivariate results vary for the fraction of industries between variable (discussed later). This is related to the relatively high observed pairwise correlation of -69.1% between this variable and across-industry similarity. Intuitively, industries that are further away likely have more industries residing between them. Our later results will show that conglomerates are more likely to operate in industry pairs that have concentrated or high value industries residing in the product space between the given pair, but not when competitive or low value industries do.

The table also shows that the average HHI variable and the within-industry similarity variable are modestly correlated at -48.7%. This result is consistent with findings in Hoberg and Phillips (2010a), and confirms that concentrated product markets generally have more product differentiation. Aside from these modest to high correlations, Table II shows that the other variables we consider have relatively low correlations. This fact, along with our very large database of 312,240 observations, indicates that multicollinearity is unlikely to be a concern in our analysis.

[**Insert Table II Here**]

Table III displays the mean values of our three key text variables for various conglomerate industry pairings. One observation is an industry pair permutation of an actual conglomerate. In Panel A, we find that conglomerates populate industries with across-industry similarity of 0.0304, which is 79% higher than the 0.017 of randomly chosen industry pairs. Conglomerates also tend to populate industries with lower than average within-industry similarity, and industries having a lower than average number of other industries between them.

[**Insert Table III Here**]

In Panel B, we report results for smaller conglomerates (two or three segments) compared to those of larger conglomerates. The table suggests that larger conglomerates tend to produce in a wider range of the product market space, as they have lower across-industry similarity. They also tend to produce in industries with more industries between them, and industries that have higher within-industry similarity. In Panel C of Table III, we observe that most conglomerates (30,525) are stable from one year to the next, although 3,259 of them reduce in size by one segment, and 600 conglomerates reduce in size by two or more segments. Analogously, 4,741 firms increase in size by one segment, and 1,644 firms increase in size by two segments.

In Panel D, we observe that vertically related conglomerates have average across-industry similarities that are close to the average for all conglomerate pairs. However the panel also shows that across-industry similarities are higher for industries having the same two digit SIC code pointing to relatedness of conglomerate chosen industry pairs. Both vertical industries and those in the same two-digit SIC code also have fewer than the average fraction of industries between them.

III Conglomerate Industry Choice

In this section we examine whether we can predict whether conglomerates produce in particular industry pairs. We test whether across-industry similarity and within-industry similarity matter for the number of conglomerate firms producing in a particular industry pair.

Table IV presents OLS regressions where each observation is a pair of three digit SIC industries in a year derived from the set of all pairings of observed SIC-3 industries in the given year in the COMPUSTAT segment tapes. The dependent variable is the **Number of Conglomerates Operating in Pair**, which is the number of conglomerates having segments in both industries associated with the given pair. Panel A displays results based on the entire sample of industry pairs. Panel B displays results for various subsamples that divide the overall sample based on the competitiveness or the valuations of industries lying between the industry pair.

[**Insert Table IV Here**]

Panel A shows that higher across-industry similarity increases the number of conglomerate firms producing in a particular industry, while average within-industry similarity decreases the conglomerate firms producing in a particular industry. Because within-industry similarity and the average HHI are moderately correlated, we examine their effects separately. The table shows that conglomerates tend to operate in more concentrated markets, ie, those with higher product differentiation and higher concentration. However, within-industry similarity matters more and we include only this variable henceforth. Panel A also shows that the fraction of industries between a given pair also matters, and its sign depends on the characteristics of the industries between.

Panels B and C show that when high value and concentrated industries are between, conglomerates operate in the pair more often. The opposite is true for competitive low value industries. This result shows how industry boundaries can be crossed and redrawn presumably by using product market synergies to lower the cost of entry into previously concentrated product markets.

Table V examines how industry characteristics influence which industry pairs are added to conglomerates in a given year. We consider raw segment additions for growing or stable conglomerates, and we also consider the SDC mergers and acquisitions database. This allows us to separately consider segments likely added through growth, or those potentially acquired in a transaction. One observation is one pair of segments in an existing conglomerate in year t. We require the conglomerate firm itself to exist in year t and year $t+1$.

The dependent variable varies by Panel. The dependent variable in Panel A is the **Number of Newly Added Conglomerate Operating Pairs**, which is the number of conglomerates having new segments in both industries associated with a given pair in a given year (where the conglomerate did not have this segment in the previous year). In Panel B, we restrict attention to new segments in conglomerates that previously had fewer segments in the previous year. Intuitively, these new segments were likely added through acquisition or organic investment. In Panel C, we restrict attention to new segments in conglomerates that were the acquirer in an acquisition in the SDC database for a transaction amounting to at least ten percent of the firm's assets. The independent variables include various product market features of the industry pair.

[**Insert Table V Here**]

The results in Panel A of Table V show that segment pairs are likely to be added if the across-industry similarity is high. This result also has the largest coefficient if the industries between two industry pairs are highly concentrated and highly valued (and the lowest coefficient when the converse is true). This result is consistent with conglomerate firms using industry links to extract product market synergies that allow them to lower the cost of entry into highly concentrated industries. We also see that conglomerate firms are more likely to add segments when the fraction of industries between the conglomerate pair is high and the average within-industry similarity is low. These findings are present especially in concentrated and highly-valued industry pairs.

The results in Panels B and C further show that conglomerate segments are more likely to be added through growth or acquisition when concentrated and highly valued industries lie between the segment pairs. In particular, conglomerate firms add such segments when the resulting industry pairs have high across-industry similarity, low within-industry similarity, and a high fraction of industries lie between the industry pair. The results are broadly consistent with conglomerates choosing to expand into industries with the potential for new differentiated products and related-industry synergy gains.

Table VI examines how industry characteristics influence which industry pairs disappear from conglomerates. Using raw segment changes and the SDC mergers and acquisitions database, we further examine segment pairs that likely were sold or closed. One observation is one pair of segments in an existing conglomerate in year t. We require the conglomerate firm itself to exist in year t and year $t + 1$.

The dependent variable again varies by Panel. In Panel A, the dependent variable is **Segment Pair Disappears**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year. In Panel B, the dependent variable is **Segment Pair Likely Sold or Closed**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate has fewer segments in year $t + 1$ relative to year t. In Panel C, the dependent variable is **Segment Pair Sold Off**, which is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate was the target of an acquisition of at least ten percent of its assets in year $t + 1$.

[**Insert Table VI Here**]

The results in Panels A and B of Table VI show that segment pairs are less likely to be sold or closed if the across-industry similarity is high. This result also has the largest coefficient if the industries between two industry pairs are highly concentrated and highly valued (and the lowest coefficient when the converse is true). Row (7) in Panel B further shows that segments are less likely to be sold or closed when there are high value concentrated industries lying between the given segment and the other segments of the conglomerate. This supports the conclusion that conglomerates see added value (such as the possibility of low cost entry and the possibility of more differentiated products) from the industries that lie between their segments.

Given these strong results regarding which industries conglomerate firms choose to operate in, we now examine how industry composition affects conglomerate valuation.

IV Conglomerate Valuation

In this section, we explore whether information in firm product descriptions can be used to construct more informative pure-play or single-segment benchmarks. We consider both product market identification and the weighting method of single-segment firms when reconstructing conglomerate benchmarks based on pure play firms. Following the existing literature, we then compare actual conglomerate valuations to the valuation of our pure play benchmarks.

A Existing Methods

Although we depart significantly from the literature in some of our conglomerate valuation methods, we begin by considering a modified algorithm based on Lang and Stulz (1994) (LS) and Berger and Ofek (1995) (BO).[8] LS and BO begin by defining a universe of candidate pure plays for each conglomerate 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 firms 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.

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

The second step following BO's framework is to compute the firm value to sales ratio for each pure play firm in each segment's universe, and then compute the median. The given segment's imputed value is then the segment's actual sales multiplied by this median ratio. Medians are used 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 the conglomerate firm is the sum of the imputed values of the given conglomerate's segments. Excess value is the natural logarithm of the conglomerate's imputed firm value divided by the conglomerate's actual firm value. This calculation can also be done using assets as an alternative to sales. A negative excess value, intuitively, suggests that the conglomerate is valued less than it might otherwise be valued if it were to operate under separate pure-play structures. We refer to this method as the "Berger+Ofek Baseline" method.

B Unconstrained Text-Based Methods

We note three key limitations of the LS and BO methods. A first is the equal treatment of all firms in a given segment's pure play universe in the 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 conglomerate. 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, if the additional pure plays are relevant, 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 conglomerate. 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 conglomerate 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 which pure plays use product vocabulary that best matches that of the conglomerate. This decomposition provides us with a set of weights, which we use 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 into parts observed in the pure play firms. Using the same notation from Section II, M_t denotes the number of unique words in the corpus, i denotes a given conglomerate being reconstructed, t denotes the year of the given conglomerate observation, and $N_{i,t}$ is the conglomerate's (M_t x 1) normalized word vector. Further suppose that the given conglomerate-year observation has $N_{it,bench}$ candidate benchmark pure play firms to use in its reconstruction. Each benchmark 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 conglomerate's observed product market vocabulary as the solution to the following least squares problem.

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

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 conglomerate. 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 conglomerate'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 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 conglomerate's TNIC industry as defined in Hoberg and Phillips (2010a). These firms have products that are similar to the conglomerate'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.

C 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 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 conglomerate 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 conglomerate'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:

$$\underset{w_{it}}{MIN}(N_{it} - BENCH_{it} \cdot w_{it})^2 \text{ such that } \; Z'_{it}w_{it} = C_{it} \tag{3}$$

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:

$$w_{it} = (BENCH_{it}'BENCH_{it})^{-1}(BENCH_{it}'N_{it} - Z_{it}\lambda), \text{ where} \tag{4}$$

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

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

D Accounting for Segment Sales

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 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 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 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. We refer to this method as the "SIC+TNIC Universe: Constrained, Segment-by-Segment" method.

V Results: Conglomerate Valuation

In this section, we first assess the quality of conglomerate reconstruction using the various reconstruction methods discussed earlier. We then briefly readdress the question of whether or not conglomerates trade at a discount. We conclude this section by testing hypotheses predicting which types of conglomerates have high or low valuations in cross section.

A Methodological Validation

Following the methodology discussion in Section IV, we examine excess valuations using five different conglomerate reconstruction methods. In particular, we consider the Berger and Ofek (1995) benchmark, and four text-based methods aimed at addressing key limitations in the BO method. Table VII displays average excess valuations, and mean squared error statistics based on these five methods. Mean excess value calculations are useful to explore if conglomerates trade at discounts (negative excess valuations) or premia (positive excess valuations), and mean squared error statistics are useful to compare the relative valuation accuracy of valuation methods. A method with a lower MSE generates excess valuations that are closer to the mean excess valuation, and are therefore more accurate. Following convention in the literature in Panels B and C, 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. In particular, more accurate valuation methods generate excess valuations outside this range less often, and thus have higher observation counts. The table reports 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 many screens to the conglomerate sample included in this part of our study. In particular, we require lagged COMPUSTAT data for our control variables, we drop firms with sales less than $20 million, firms with zero assets, and firms for which summed segment sales disagrees with the overall firm's sales by more than 1%. We also 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. In Panel A of VII we include all observations. In Panel B, following the convention in the literature, we restrict attention to excess values less than 400% and greater than 25% (screen applied separately for each method). In Panel C, we omit all firm-years for a conglomerate in which 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).

[**Insert Table VII Here**]

Panel A of Table VII 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, just changing the weighting of single segment firms alone is partially but not fully responsible for our ability to explain the discount. Row 3 of Panel A expands the universe to include TNIC pure play rivals of the conglomerate. This expansion reduces the discount to 4.6%. Finally, using the five key accounting characteristics in Row 4 reduces the discount to 1.2%. In row 5 of Panel A, we see that further constraining the weights to match segment-specific sales ratios increases the discount to just 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 conglomerate pricing. When based on sales, the mean squared error in row 4 of .320 is 32.4% smaller than the mean squared error of .474 associated with the Berger and Ofek benchmark. When based on assets, this improvement is 27.7%.

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

We conclude that improving conglomerate benchmarks alone can explain the previously reported conglomerate discount, and 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 to the conglomerate itself. Our results therefore do not support the conclusion that conglomerate firms trade at discounts. These findings are in line with other recent studies that draw the same conclusion using other methods (see Campa and Kedia (2002), Villalonga (2004b), and Graham, Lemmon, and Wolf (2002)).

In Table VIII, we assess whether conglomerates reconstructed using the various methods discussed above have similar characteristics as the conglomerates themselves. As the objective of these methods is to rebuild an identical replica of what the conglomerate would look like under a non-conglomerate structure, better benchmarks should match the conglomerate along more dimensions. For example, they should have similar sales growth, should be equally as mature, should be as profitable, and they should have similar investment intensities.

To address this question, we first compute implied characteristic values using the same methods used to compute imputed valuations in the excess value valuations. For example, the implied Sales Growth of a Berger and Ofek (baseline) benchmark is computed as the sales weighted average of the segment-by-segment computed median sales growth of the pure plays in each segment's three digit SIC industry. For a text-based benchmark, the weighted median sales growth is the implied sales growth of the conglomerate.

[Insert Table VIII Here]

Table VIII reports correlations between the actual conglomerate characteristic and the implied characteristic for each characteristic noted in the first column using each valuation method noted in the remaining columns. Comparing correlations between the single-segment constructed benchmarks and the actual conglomerate firms using the Berger and Ofek benchmark to the text-based benchmarks reveals that the text-based benchmarks strongly outperform the Berger and Ofek baseline in terms of matching characteristics. The simplest text-based methods that do not constrain accounting characteristics (columns two and three) have higher correlations than the Berger and Ofek constructed conglomerate benchmark. For example, the 28.9% correlation between the OI/Assets of the actual conglomerate and the Berger and Ofek benchmark increases dramatically to (35.7% to 42.1%) even using unconstrained text-based weights. As indicated in the methodology section, the text-based weights are purely a function of the vocabulary used by the pure plays and the conglomerate, and are not mechanistically related to the accounting numbers that these methods are better able to match. 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 use five key accounting characteristics to better fit each conglomerate's mapping, it is not surprising that these characteristic correlations are higher. We conclude that text-based measures offer substantial improvements over existing methods.

It is also natural to ask which type of pure play firms are weighted more than others when reconstructing conglomerates and giving differential weights to component single-segment firms. Panel A of Table IX explores this question and displays average characteristics for firms assigned weights in the highest and lowest quartile using the text-based conglomerate benchmarks. Panel B further examines how the weights on the single-segment benchmark firms vary when we examine how well the vocabularies of the pure plays fit the vocabulary of the conglomerate (we define the "Difficulty of Pure Plays to Replicate the Conglomerate" as one minus the R^2 from the vocabulary decomposition).

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 pure plays (SIC and TNIC peers) to reconstruct the conglomerates. In the second three columns in Table IX, we repeat the same exercise using the "SIC+TNIC universe (constrained)" method, which also holds fixed key accounting variables as discussed earlier.

[**Insert Table IX 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 inappropriate conclusion that conglomerates are undervalued. Our results in the next section formally confirm this conjecture.

Panel B Table IX provides a similar comparison of characteristics but splits the sample based on which firms have above or below median "Difficulty of Pure Plays to Replicate" measures (1 - R^2 from the textual decomposition regression in equation (3)). A conglomerate that is difficult to replicate has an R^2 of zero and a "Difficulty of Pure Plays to Replicate" value of 1. This constrained regression is run once per conglomerate-year, as this provides us with the weights used to construct the excess values year by year as discussed in the previous section. This same calculation thus provides one difficulty to replicate statistic for each conglomerate in each year..

Panel B shows that conglomerates with concentrated industries and high-value industries between their segments have the sharpest correlations between difficulty to replicate and the fractions of industries lying between the pair. This finding adds to our earlier evidence that conglomerates spanning these high value industries generate more product market synergies, as these same synergies likely explain why the pure plays cannot easily replicate these conglomerates.

B Determinants of Conglomerate Valuations

In this section, we examine whether conglomerate valuations vary in cross section. As discussed in our hypotheses section (Section I), we focus on examining whether conglomerates that are harder to replicate and face less competition have higher valuations relative to our pure-play based benchmarks. To explore this question, we regress conglomerate excess valuation on the text-based variables that capture these factors.

We regress conglomerate-year excess valuations on the "Difficulty of Pure Plays to Replicate" variable, the fraction of industries between the conglomerate industry segments, concentration measures, and across and within-industry similarity measures. We also include controls for document length, vertical relatedness, and accounting variables used in the existing literature.

The "Difficulty of Pure Plays to Replicate" variable captures how easily the conglomerate can be reconstructed using the set of pure play firms that exist in its markets. The intuition underlying this calculation is that a conglomerate that is more difficult to replicate is likely more differentiated, and hence faces less of a competitive threat. For example, product market synergies created through its conglomerate structure cannot be easily raided by newly formed conglomerates based on existing pure plays.

The fraction of industries between the conglomerate industry segment captures the potential gains a conglomerate can reap by expanding its product offerings into these between markets. Average across-industry similarity captures the potential for synergies between industries. Within industry similarity and concentration capture the degree of competition that a conglomerate firm faces in its segment markets.

[**Insert Table X Here**]

Table X displays the results of OLS panel data regressions in which one observation is one conglomerate in one year, and the dependent variable is its excess valuation using the constrained text-based valuation method (Panel A) and the Berger and Ofek (1995) valuation method (Panel B). t-statistics are shown in parentheses, and standard errors are adjusted for clustering by firm.

Our first key finding is that the difficulty of pure plays to replicate variable is positive and highly statistically significant in both panels. Conglomerates that are harder to replicate have high valuations relative to pure play benchmarks. As this variable captures the uniqueness of the conglomerate's products relative to the pure play benchmarks, one would not expect its affect on valuation to be negated out in the difference as was the case for the average HHI variable. This finding, which is robust at the 1% level of significance in all rows, is consistent with these firms earning higher rents due to the inability of other firms to replicate the product market synergies they enjoy from the conglomerate structure. Our control variables indicate that conglomerates are also valued more when they have more investment (R+D and Capital Expenditures), when they are more profitable, and when they are larger. Conglomerates are also less valuable when their segments are vertically related.

We also find that the reported R^2s are higher in Panel B than in Panel A. 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 in Panel A than in Panel B. The table also shows that the level of significance of our key variable, difficulty of pure plays to replicate, is quite similar in both panels, and thus it is robust to changes in the quality of the match.

The table also shows that average within-segment similarity and the average concentration ratio (Conglomerate Average Concentration) are not significantly related to excess valuations in the full model in Row (6), although they are both significant when the Difficulty to Replicate variable is excluded in rows (2) and (4). These results are not surprising because the pure play firms used to construct excess valuations enjoy the same level of concentration on average as the conglomerate. Stated differently, the differencing used to construct excess valuations should negate at least some of the effect of industry characteristics such as these (although it would have no effect on the difficulty to replicate variable as that is a unique property of the observed conglomerate).

Table XI displays 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 X with the exception of the difficulty to replicate variable. The table shows that raw 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%). We conclude that the impact of a conglomerate's difficulty to reconstruct is meaningful, and that conglomerates that are more difficult to replicate trade at modest premia relative to their pure play benchmarks.

[**Insert Table XI Here**]

Our last table examines conglomerate excess valuation for subsamples based on which industries conglomerate firms operate within. Panel A of Table XII considers conglomerates with high-value industries (above median) between their segment pairs, and Panel B examines conglomerate excess valuations for conglomerates with low-value industries (below median) between their segment pairs.

[**Insert Table XII Here**]

The results in Panel A, as compared to Panel B, show that the "Difficulty of Pure Plays to Replicate" variable is significant in both panels but significantly larger for high-value industries. We can also see that the fraction of industries between is positive and significant for high-value industries but insignificant for conglomerates with low-value industries between their industry pairs. As in Table X, the difficulty to replicate variable subsumes other product market variables, as this is a comprehensive measure of how differentiated the conglomerate is from potential rivals in all of its markets.

Overall these results, combined with those in Table X, are consistent with conglomerates having higher valuations when their products are difficult to replicate with the best possible single-segment peers. This suggests that such conglomerates extract more value when the conglomerate structure allows product market synergies that can differentiate product offerings, and when such gains cannot be replicated by single-segment rival firms.

VI Conclusions

We use text-based computational analysis of conglomerate and pure-play firm business descriptions from 10-Ks filed with the SEC to examine in which industries conglomerates are most likely to operate and to understand cross-sectional differences in conglomerate valuations. We find that conglomerate firms are more likely to operate in industry pairs that are closer together in the product space, in industry pairs that have highly valued product spaces "between" them, and in industries with lower within-industry product similarity. These findings are consistent with firms using the multiple-industry structure to try to take advantage of economies of scope and potential product synergies across markets to create differentiated products that are not easily replicated by single-segment rivals.

We examine the cross-sectional valuation differences for conglomerate firms using text-based analysis and vocabulary decompositions to redefine conglomerate benchmarks constructed from single-segment firms using product-word vocabulary matching. Our benchmarks generate dramatic improvements in valuation accuracy, and in characteristic matches relative to existing methods. Our benchmarks also provide a measure of how differentiated the conglomerate firm's products are relative to single-segment matched firms.

We find that on average conglomerates do not trade at a discount relative to text-matched single-segment firms. More importantly, this average effect masks important cross-sectional variation. Conglomerate firms that are more difficult to reconstruct and are likely to have more differentiated products relative to single-segment firms trade at stock-market premiums, while conglomerate firms that are easy to reconstruct or "span" using single-segment firms are more likely to trade at stock-market discounts. These findings are consistent with higher valued conglomerate firms extracting product market synergies that enhance product differentiation, and with successful conglomerate production lowering the cost of entry into valuable industries between industry pairs.

References

Ahern, Kenneth, and Jarrad Harford, 2011, The importance of industry links in merger waves, University of Michigan and University of Washington Working Paper.

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).

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.

Denis, David, Diane Denis, and Atulya Sarin, 1997, Agency problems, equity ownership and corporate diversification, *Journal of Finance* 52, 135–160.

Fan, Joseph, and Vidhan Goyal, 2006, On the patterns and wealth effects of vertical mergers, *Journal of Business* 79, 877–902.

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

Hay, D.A., 1976, Sequential entry and entry-deterring strategies in spatial competition, *Oxford Economic Papers* 28, 240–257.

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

Hund, John, Donald Monk, and Sheri Tice, 2010, Uncertainty about average profitability and the diversification discount, *Journal of Financial Economics* 96, 463–484.

Kedia, Simi, Abraham Ravid, and Vicente Pons, 2008, Vertical mergers and the market valuation of the benefits of vertical integration, Rutgers Business School Working Paper.

Khanna, Naveen, and Sheri Tice, 2001, The bright side of internal capital markets, *Journal of Finance* 56, 1489–1528.

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.

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

Maksimovic, Vojislav, and Gordon Phillips, 2002, Do conglomerate firms allocate resources inefficiently across industries? theory and evidence, *Journal of Finance* 57, 721–767.

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

Matsusaka, John, 2001, Corporate diversification, value maximization, and organizational capabilities, *Journal of Business* 74, 409–431.

Panzar, J., and R. Willig, 1981, Economies of scope, *American Economic Review* 71, 268–272.

Rajan, Raghuram G., Henri Servaes, and Luigi Zingales, 2000, The cost of diversity: the diversification discount and inefficient investment, *Journal of Finance* 55, 35–80.

Rhodes-Kropf, Matthew, and David Robinson, 2008, The market for mergers and the boundaries of the firm, *Journal of Finance* 63, 1169–1211.

Scharfstein, David, and Jeremy Stein, 2000, The dark side of internal capital markets: Segment rent seeking and inefficient investments, *Journal of Finance* 55, 2537–2564.

Schoar, Antoinette, 2002, The effect of diversification on firm productivity, *Journal of Finance* 62, 2379–2403.

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.

Shin, Hyun-Han, and Rene M. Stulz, 1998, Are internal capital markets efficient?, *Quarterly Journal of Economics* 113, 531–552.

Stein, Jeremy, 1997, Internal capital markets and the competition for corporate resources, *Journal of Finance* 52, 111–133.

———, 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: Summary Statistics

Summary statistics are reported for our sample of conglomerate and pure play firms (Panel A), Industry Pairs and conglomerates (Panel B), and Conglomerate Segment Pairs (Panel C) for our sample from 1996 to 2008.

Variable	Mean	Std. Dev.	Minimum	Median	Maximum
Panel A: Conglomerates (15,373 obs) and Pure-Plays (56,491 obs)					
Firm Value (Conglomerates)	12430	48462	0.483	1228	1036340
Firm Value (Pure-Plays)	2450	18863	0.003	215.	1038648
TNIC HHI (Conglomerates)	0.140	0.219	0.006	0.059	1.000
TNIC HHI (Pure-Plays)	0.111	0.153	0.006	0.058	1.000
Panel B: Industry Pairs (312,240 obs) and Conglomerates (15,373 obs)					
Number of Conglomerates Operating in Pair (Ind. Pairs)	0.147	0.855	0.0	0.0	57.0
Across Industry Similarity (Ind. Pairs)	0.017	0.010	0.000	0.014	0.169
Across Industry Similarity (Conglomerates)	0.032	0.019	0.000	0.025	0.138
Fraction of Industries Between Pair (Ind. Pairs)	0.325	0.257	0.000	0.267	0.992
Fraction of Industries Between Pair (Conglomerates)	0.097	0.133	0.000	0.042	0.992
Within Industry Similarity (Ind. Pairs)	0.086	0.038	0.000	0.081	0.433
Within Industry Similarity (Conglomerates)	0.073	0.030	0.010	0.066	0.188
Same 2-digit SIC Dummy (Ind. Pairs)	0.018	0.133	0.000	0.000	1.000
Same 2-digit SIC Dummy (Conglomerates)	0.228	0.371	0.000	0.000	1.000
Vertical Relatedness (Ind. Pairs)	0.003	0.014	0.000	0.000	0.536
Vertical Relatedness (Conglomerates)	0.027	0.066	0.000	0.006	0.536

Variable	Obs	Percentage	Std. Dev.
Panel C: Change in Conglomerate Segment Pair Variables (32,181 obs)			
Segment Pair Disappears	4,566	14.2%	34.9%
Segment Pair Likely Sold or Closed	3,415	10.6%	30.8%
Segment Pair Likely Sold Off	330	1.0%	10.1%

Table II: Pearson Correlation Coefficients

Pearson Correlation Coefficients are reported for our sample of 312,240 observations of three digit SIC industry pairs from 1996 to 2008.

Row	Variable	Number of Operating Conglom. Pairs	Across Industry Similarity	Zero Industries Between Dummy	Fraction of Industries Between	Within Industry Similarity	Average HHI	Pair Likelihood if Random	Same 2-digit SIC Dummy
				Correlation Coefficients					
(1)	Across Industry Similarity	0.229							
(2)	Zero Industries Between Dummy	0.160	0.446						
(3)	Fraction of Industries Between Pair	-0.132	-0.691	-0.137					
(4)	Within Industry Similarity	-0.044	0.184	0.058	-0.092				
(5)	Average HHI	-0.011	-0.176	-0.042	0.088	-0.487			
(6)	Pair Likelihood if Random	0.144	-0.009	0.020	-0.002	-0.020	-0.031		
(7)	Same 2-digit SIC Dummy	0.231	0.315	0.200	-0.135	-0.030	0.020	0.012	
(8)	Vertical Relatedness	0.200	0.165	0.078	-0.124	-0.049	0.055	0.028	0.155

Table III: Conglomerate Summary

Summary statistics for various industry pairs from 1996 to 2008. Panel A compares observed conglomerate industry pairs to randomly drawn industry pairs. Panel B displays observed conglomerate industry pairs for conglomerates of varying size. Panel C displays conglomerate industry pairs for conglomerates that are growing, stable, or shrinking, as noted in the first column. Panel D displays conglomerate industry pairs for vertically integrated segments and for segments that are in the same two-digit SIC code.

Sub Sample	Across Industry Similarity	Within Industry Similarity	Average HHI	Fraction of Industries Between	# Obs.
Panel A: Overall					
All Conglomerates	0.0296	0.0768	0.1150	0.1293	40,769
Randomly Drawn SIC-3 Industries	0.0167	0.0862	0.1183	0.3255	312,240
Panel B: By Conglomerate Size					
2 Segments	0.0341	0.0738	0.1192	0.0867	6,365
3 Segments	0.0311	0.0750	0.1164	0.1132	11,672
4-5 Segments	0.0289	0.0786	0.1130	0.1366	15,794
6+ Segments	0.0247	0.0785	0.1133	0.1790	6,938
Panel C: Shrinking, Stable, and Growing Conglomerates					
Shrink by 2+ Segments	0.0268	0.0788	0.1097	0.1490	600
Shrink by 1 Segment	0.0295	0.0779	0.1119	0.1296	3,259
Stable Conglomerate	0.0301	0.0769	0.1160	0.1260	30,525
Add 1 Segment	0.0282	0.0760	0.1117	0.1414	4,741
Add 2+ Segments	0.0262	0.0739	0.1135	0.1485	1,644
Panel D: Vertical and Same SIC-2 Conglomerates					
Vertically Related Segments	0.0319	0.0717	0.1212	0.0739	15,007
Same SIC-2 Segments	0.0471	0.0829	0.1085	0.0291	8,015

Table IV: Where Conglomerates Exist

OLS regressions with year fixed effects and standard errors clustered by year for our sample of 312,240 industry pairs from 1996 to 2008. One observation is one pair of three digit SIC industries in a year derived from the set of all permutations of feasible pairings. The dependent variable is the number of conglomerates operating in the given industry pair. Panel A displays results based on the entire sample. Panels B and C display results for subsamples based on the competitiveness and valuations of industries lying between the given industry pair.

Row	Sample	Across Industry Similarity	Fraction of Industries Between Pair	Zero Industries Between	Avg. Within Industry Similarity	Aver- age HHI	Pair Likeli- hood if Random	Same 2-digit SIC Code	Vertical Relat- edness	# Obs. / RSQ
				Panel A: Full Sample						
(1)	All Industry Pairs	14.060	0.060	0.410	-1.347		0.084	0.943	8.669	312,240
		(19.98)	(4.85)	(6.20)	(-13.90)		(9.32)	(18.94)	(7.00)	0.128
(2)	All Industry Pairs	12.809	0.045	0.423		0.181	0.085	0.973	8.869	312,240
		(18.94)	(3.45)	(6.40)		(4.05)	(9.42)	(19.14)	(7.13)	0.125
				Panel B: Univariate Subsamples						
(3)	Concentrated Industry Pairs	27.374	0.249		-1.034		0.086	0.638	3.715	154,324
		(11.15)	(6.32)		(-18.74)		(6.65)	(8.57)	(7.56)	0.110
(4)	Competitive Industry Pairs	12.730	-0.050		-1.625		0.076	1.044	8.033	154,321
		(16.76)	(-1.89)		(-11.18)		(6.07)	(20.31)	(7.67)	0.103
(5)	High Firm Value Industry Pairs	21.110	0.190		-1.260		0.063	1.199	5.695	154,326
		(12.96)	(5.51)		(-11.80)		(6.19)	(19.43)	(4.14)	0.100
(6)	Low Firm Value Industry Pairs	11.380	-0.010		-1.453		0.120	0.743	8.491	154,319
		(15.52)	(-1.43)		(-10.86)		(5.91)	(12.26)	(12.13)	0.124
				Panel C: Bivariate Subsamples						
(7)	Concentrated and High Value Pairs	38.414	0.425		-0.865		0.066	0.779	3.207	65,904
		(6.04)	(4.29)		(-12.71)		(3.99)	(5.53)	(4.11)	0.113
(8)	Competitive and High Value Pairs	19.416	0.160		-1.534		0.062	1.294	6.165	88,422
		(12.14)	(3.11)		(-10.37)		(5.93)	(16.84)	(3.93)	0.097
(9)	Concentrated and Low Value Pairs	22.061	0.146		-1.153		0.113	0.595	3.937	88,420
		(8.88)	(4.20)		(-13.12)		(4.06)	(7.94)	(7.39)	0.114
(10)	Competitive and Low Value Pairs	8.544	-0.258		-1.813		0.124	0.817	10.600	65,899
		(9.38)	(-14.40)		(-9.73)		(4.70)	(13.45)	(8.77)	0.127

Table V: New Conglomerate Segments

OLS regressions with year fixed effects and standard errors clustered by year. The dependent variable is the number of new conglomerate segments in each three-digit SIC code pair in the given year. Panel A counts the number of new conglomerates operating in both industries of an industry pair. Panel B restricts attention to new segments from conglomerates that had fewer segments in the previous year. Panel C restricts attention to new segments of conglomerates that were the acquirer in a transaction amounting to at least ten percent of the firm's assets.

Row	Sample	Across Industry Similarity	Fraction Industries Between Pair	Avg. Within Industry Simil.	Pair Likeli-hood if Random	Same 2-digit SIC Code	Vertical Relat-edness	Obs. /RSQ
				Panel A: Dep. Var = New Segment Pairs				
(1)	All Industry Pairs	2.409	0.010	-0.268	0.017	0.124	0.756	312,240
		(5.56)	(3.23)	(-4.41)	(3.86)	(4.69)	(2.61)	0.052
(2)	Concen. + High Value	6.016	0.065	-0.160	0.016	0.130	0.519	65,904
		(4.13)	(3.11)	(-4.78)	(2.50)	(2.61)	(1.96)	0.051
(3)	Concen. + Low Value	2.348	-0.009	-0.337	0.014	0.155	0.541	88,422
		(7.02)	(-0.84)	(-4.49)	(3.42)	(5.85)	(2.08)	0.046
(4)	Compet. + High Value	3.051	0.021	-0.176	0.019	0.076	0.758	88,420
		(4.15)	(2.63)	(-4.39)	(3.56)	(2.58)	(2.91)	0.038
(5)	Compet. + Low Value	1.416	-0.036	-0.341	0.024	0.111	0.829	65,899
		(3.16)	(-3.69)	(-3.75)	(2.77)	(3.60)	(2.45)	0.049
				Panel B: Dep. Var = New Segment Pairs Likely Obtained through Growth				
(6)	All Industry Pairs	1.994	0.009	-0.213	0.014	0.101	0.614	312,240
		(4.92)	(3.84)	(-3.99)	(3.61)	(4.29)	(2.37)	0.048
(7)	Concen. + High Value	4.249	0.042	-0.125	0.014	0.130	0.415	65,904
		(4.36)	(3.37)	(-4.06)	(2.30)	(2.61)	(1.60)	0.046
(8)	Concen. + Low Value	1.826	-0.010	-0.272	0.012	0.125	0.479	88,422
		(5.26)	(-1.01)	(-4.14)	(3.25)	(5.74)	(2.18)	0.043
(9)	Compet. + High Value	2.384	0.016	-0.134	0.015	0.062	0.602	88,420
		(3.63)	(2.29)	(-4.20)	(3.05)	(2.50)	(2.76)	0.033
(10)	Compet. + Low Value	1.260	-0.028	-0.272	0.022	0.087	0.640	65,899
		(3.17)	(-3.11)	(-3.31)	(2.67)	(3.06)	(2.07)	0.046
				Panel C: Dep. Var = New Segment Pairs Linked to SDC Acquisitions				
(11)	All Industry Pairs	0.239	0.002	-0.019	0.001	0.004	0.073	312,240
		(3.99)	(2.09)	(-3.13)	(2.43)	(1.79)	(2.31)	0.007
(12)	Concen. + High Value	0.605	0.007	-0.009	0.001	-0.001	0.036	65,904
		(2.44)	(2.22)	(-3.20)	(1.81)	(-0.35)	(1.24)	0.005
(13)	Concen. + Low Value	0.260	0.001	-0.021	0.001	0.006	0.046	88,422
		(4.19)	(0.46)	(-2.88)	(2.36)	(1.78)	(1.53)	0.007
(14)	Compet. + High Value	0.278	0.002	-0.012	0.001	0.003	0.040	88,420
		(2.41)	(1.32)	(-3.27)	(3.49)	(1.07)	(2.27)	0.004
(15)	Compet. + Low Value	0.176	-0.002	-0.033	0.001	0.005	0.091	65,899
		(2.18)	(-0.72)	(-2.91)	(1.52)	(2.03)	(1.35)	0.007

Table VI: Which Segments Exit

Logistic regressions with year fixed effects and standard errors clustered by year for our sample of 32,181 observed conglomerate industry pairs from 1997 to 2008. In Panel A, the dependent variable is a dummy equal to one if the given segment pair does not exist in the given conglomerate's structure in the following year. In Panel B, the dependent variable is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate's number of segments has declined. In Panel C, the dependent variable is a dummy equal to one if the given pair does not exist in the conglomerate's structure in the following year, and the conglomerate was the target of an acquisition of at least ten percent of its assets.

Row	Sample	Across Industry Similarity	Fraction Industries Between Pair	Avg. Within Industry Simil.	Pair Likeli-hood if Random	Same 2-digit SIC Code	Vertical Relat-edness	Obs. /RSQ
	Panel A: Dep. Var = Segment Pair Disappears							
(1)	All Pairs	-6.557	0.282	0.362	0.004	-0.043	-1.724	32,181
		(-2.94)	(1.75)	(0.32)	(0.67)	(-1.23)	(-4.13)	0.015
(2)	Concen. + High Value	-17.662	-0.047	-2.029	0.013	-0.146	-0.213	7,387
		(-2.83)	(-0.15)	(-1.20)	(1.53)	(-1.16)	(-0.17)	0.015
(3)	Compet. + High Value	-10.653	-0.120	0.498	-0.004	-0.135	-2.493	6,976
		(-3.45)	(-0.13)	(0.31)	(-0.43)	(-2.15)	(-2.27)	0.024
(4)	Concen. + Low Value	-15.653	0.024	0.259	0.006	-0.131	1.076	8,706
		(-2.40)	(0.09)	(0.13)	(0.51)	(-1.50)	(0.60)	0.011
(5)	Compet. + Low Value	-3.574	2.919	0.896	-0.014	0.005	-1.192	5,636
		(-0.78)	(1.33)	(0.68)	(-0.69)	(0.09)	(-2.17)	0.013
	Panel B: Dep. Var = Segment Pair Likely Sold or Closed							
(6)	All Pairs	-8.521	-0.004	1.166	0.008	-0.137	-1.692	32,181
		(-3.14)	(-0.02)	(0.98)	(1.17)	(-2.37)	(-3.48)	0.009
(7)	Concen. + High Value	-22.566	-0.507	-1.110	0.015	-0.424	0.124	7,387
		(-3.18)	(-2.03)	(-0.50)	(1.67)	(-2.43)	(0.19)	0.011
(8)	Compet. + High Value	-14.508	-1.009	0.065	-0.003	-0.221	-2.499	6,976
		(-2.46)	(-0.99)	(0.04)	(-0.24)	(-1.91)	(-1.99)	0.016
(9)	Concen. + Low Value	-14.934	-0.262	1.662	0.009	-0.198	1.336	8,706
		(-2.09)	(-0.94)	(1.01)	(0.83)	(-2.92)	(0.91)	0.007
(10)	Compet. + Low Value	-6.072	1.303	2.356	-0.004	-0.082	-1.312	5,636
		(-0.88)	(0.67)	(1.22)	(-0.20)	(-0.91)	(-1.55)	0.007
	Panel C: Dep. Var = Segment Pair Sold Off							
(11)	All Pairs	-2.326	0.186	-0.229	0.004	0.085	0.120	32,181
		(-0.25)	(0.54)	(-0.07)	(0.29)	(0.31)	(0.12)	0.004
(12)	Concen. + High Value	-27.734	-0.305	-8.540	0.021	0.655	2.888	7,387
		(-1.61)	(-0.73)	(-1.18)	(2.51)	(1.26)	(0.71)	0.009
(13)	Compet. + High Value	11.600	0.115	1.604	0.010	-0.385	-2.983	6,976
		(0.84)	(0.04)	(0.37)	(0.33)	(-0.97)	(-0.80)	0.006
(14)	Concen. + Low Value	2.023	-0.258	3.912	-0.009	0.046	0.726	8,706
		(0.15)	(-0.24)	(0.77)	(-0.26)	(0.15)	(0.21)	0.007
(15)	Compet. + Low Value	-0.911	10.004	-0.353	-0.552	0.147	0.945	5,636
		(-0.04)	(2.43)	(-0.07)	(-1.15)	(0.39)	(0.39)	0.004

Table VII: Quality of Excess Valuation Calculations Across Methods

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 VIII: Characteristic Correlations (Conglomerate vs. Benchmark)

The table displays Pearson Correlation coefficients between actual conglomerate characteristics and implied characteristics using several different conglomerate benchmark methods as noted in the column headers.

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)
			Correlation Coefficients			
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	TNIC HHI	0.325	0.430	0.535	0.516	0.387
9	Log Age	0.268	0.298	0.436	0.924	0.731

Table IX: Which Pure Plays Match with Conglomerates?

The table displays summary statistics for pure play firms assigned above median weights versus below median weights in conglomerate benchmarks (Panel A), and conglomerates with above median and below median difficulty to replicate using pure plays (Panel B).

Panel A: 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
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	TNIC HHI	0.08	0.07	-12.58	0.08	0.07	-9.48
9	Firm Age	25.32	29.19	18.63	24.06	30.32	29.14

Panel B: Difficulty of Pure Plays to Replicate vs Characteristics

Row	Variable	Lowest Difficulty to Replicate	Highest Difficulty to Replicate	t-statistic of Difference	Lowest Difficulty to Replicate	Highest Difficulty to Replicate	t-statistic of Difference
		Conglomerates with Concentrated Industries Between			*Conglomerates with Competitive Industries Between*		
10	Fraction of Ind. Between	0.142	0.224	5.951	0.046	0.054	1.940
11	Across Ind. Similarity	0.023	0.019	-7.193	0.041	0.032	-7.913
12	Within Ind. Similarity	0.063	0.047	-8.153	0.091	0.063	-12.133
13	TNIC HHI	0.053	0.082	10.054	0.029	0.059	17.454
		Conglomerates with High Value Industries Between			*Conglomerates with Low Value Industries Between*		
14	Fraction of Ind. Between	0.079	0.203	10.213	0.071	0.131	6.561
15	Across Ind. Similarity	0.035	0.023	-11.465	0.034	0.024	-9.119
16	Within Ind. Similarity	0.084	0.058	-11.289	0.073	0.050	-10.204
17	TNIC HHI	0.032	0.068	15.182	0.047	0.075	11.201

Table X: Conglomerate Excess Valuations

OLS regressions with time fixed effects and standard errors clustered by firm. One observation is one conglomerate from 1997 to 2008. The dependent variable is the conglomerate's excess valuation 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 VII.

Row	Difficulty of Pure Plays to Replicate	Fraction of Indust. Between	Across Segment Similarity	Within Segment Similarity	Conglom. Average Concen-tration	Log Document Length	Vertical Relat-edness	R&D/ Sales	CAPX/ Sales	OI/ Sales	Log Assets	# Obs. / RSQ
				Panel A: Excess Value (Text-based Constrained Valuation Model)								
(1)	0.406		.		.	-0.026	-0.338	1.287	0.467	0.603	0.044	9,201
	(7.59)					(-1.35)	(-1.44)	(6.42)	(7.08)	(8.58)	(9.07)	0.100
(2)		0.120		.		-0.075	-0.282	1.212	0.453	0.600	0.039	9,201
		(1.92)				(-3.97)	(-1.19)	(5.97)	(6.97)	(8.50)	(7.95)	0.088
(3)		.	-0.888		.	-0.073	-0.259	1.147	0.453	0.594	0.040	9,201
			(-1.98)			(-3.88)	(-1.08)	(5.53)	(6.98)	(8.41)	(8.07)	0.088
(4)		.		-0.616		-0.068	-0.309	1.102	0.459	0.600	0.040	9,201
				(-2.48)		(-3.58)	(-1.31)	(5.29)	(6.99)	(8.48)	(8.20)	0.089
(5)		.		.	0.508	-0.067	-0.355	1.144	0.463	0.597	0.040	9,201
					(2.06)	(-3.48)	(-1.51)	(5.58)	(7.03)	(8.47)	(8.17)	0.088
(6)	0.389	0.060	0.156	-0.229	0.009	-0.023	-0.286	1.263	0.473	0.604	0.045	9,201
	(7.04)	(0.87)	(0.28)	(-0.80)	(0.04)	(-1.14)	(-1.18)	(6.11)	(7.09)	(8.59)	(9.14)	0.101
				Panel B: Excess Value (Berger + Ofek Valuation Model)								
(7)	0.549		.		.	0.056	0.003	2.454	0.688	1.113	0.063	8,951
	(8.23)					(2.41)	(0.01)	(10.55)	(7.58)	(12.36)	(10.84)	0.198
(8)		0.099		.		-0.012	0.022	2.314	0.659	1.105	0.055	8,951
		(1.31)				(-0.54)	(0.08)	(9.77)	(7.32)	(12.18)	(9.49)	0.179
(9)		.	-1.273		.	-0.007	0.116	2.222	0.664	1.096	0.056	8,951
			(-2.37)			(-0.29)	(0.39)	(9.26)	(7.38)	(12.11)	(9.68)	0.180
(10)		.		-1.130		0.005	0.077	2.122	0.681	1.106	0.058	8,951
				(-3.94)		(0.21)	(0.26)	(8.85)	(7.55)	(12.25)	(9.88)	0.182
(11)		.		.	0.348	-0.008	-0.041	2.264	0.665	1.102	0.056	8,951
					(1.10)	(-0.33)	(-0.14)	(9.53)	(7.36)	(12.18)	(9.52)	0.179
(12)	0.542	-0.007	0.045	-0.720	-0.473	0.057	0.054	2.392	0.688	1.118	0.063	8,951
	(7.96)	(-0.08)	(0.07)	(-2.19)	(-1.52)	(2.33)	(0.18)	(9.96)	(7.57)	(12.43)	(10.78)	0.199

Table XI: Economic Magnitudes and Excess Valuation

This table displays average excess valuations for quintiles based on the difficulty of pure plays to replicate. 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 X 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			
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

Table XII: Conglomerate Excess Valuations (Various Subsamples)

OLS regressions with time fixed effects and standard errors clustered by firm. One observation is one conglomerate from 1997 to 2008. The dependent variable is the conglomerate's excess valuation using the "HP: SIC+TNIC Universe: Constrained" model. We consider various subsamples as noted in the panel headers.

Row	Difficulty of Pure Plays to Replicate	Fraction of Indust. Between	Across Segment Similarity	Within Segment Similarity	Conglom. Average Concen-tration	Log Document Length	Vertical Relat-edness	R&D/ Sales	CAPX/ Sales	OI/ Sales	Log Assets	# Obs. / RSQ
					Panel A: Excess Value (Conglomerates with High Value Industries Between)							
(1)		0.183		.		-0.075	0.200	1.391	0.486	0.561	0.041	4,178
		(2.34)				(-2.97)	(0.60)	(4.70)	(6.49)	(6.39)	(6.20)	0.097
(2)	0.473	0.098	-0.084	-0.396	-0.120	-0.015	0.238	1.426	0.507	0.567	0.047	4,178
	(5.08)	(1.13)	(-0.09)	(-0.96)	(-0.32)	(-0.55)	(0.68)	(4.70)	(6.61)	(6.44)	(7.12)	0.111
					Panel B: Excess Value (Conglomerates with Low Value Industries Between)							
(3)		0.062		.		-0.073	-0.520	0.806	0.455	0.664	0.036	4,172
		(0.74)				(-2.96)	(-1.67)	(2.70)	(3.86)	(5.36)	(5.80)	0.079
(4)	0.323	-0.107	-1.726	-0.089	-0.214	-0.036	-0.544	0.784	0.455	0.640	0.043	4,172
	(5.11)	(-1.10)	(-1.83)	(-0.23)	(-0.74)	(-1.40)	(-1.73)	(2.61)	(3.79)	(5.20)	(6.79)	0.092