Looking for heterogeneous firms : sources and implications for financial statement users

Thèse

Baptiste Colas

Doctorat en sciences de l'administration
Philosophiæ doctor (Ph. D.)

Québec, Canada

Looking for heterogeneous firms: sources and implications for financial statement users

Thèse

Baptiste Colas

Sous la direction de :

Carl Brousseau, directeur de recherche

Résumé

Cette thèse s'intéresse à la sélection de comparables dans le contexte de la comptabilité financière. Dans ce contexte, l'analyse de firmes se fait de façon relative, en comparaison avec d'autres firmes semblables — les « comparables ». Ainsi, il est nécessaire de former des groupes homogènes de firmes à ces fins. L'utilisation des classifications d'industries est la méthode privilégiée, car elle permet de grouper les firmes sur des critères objectifs et en lien avec le cœur de l'activité des firmes. Elles présentent l'avantage d'être très largement disponible, et très simples à utiliser. Dans cette thèse l'objectif principal est d'identifier des sources d'hétérogénéité intra-industrie, et d'examiner leurs conséquences à plusieurs niveaux. J'utilise trois approches différentes pour atteindre cet objectif.

Dans un premier temps, l'objectif est de proposer une utilisation plus complète des classifications d'industries. Ainsi, j'utilise exclusivement les classifications d'industries pour identifier une source d'hétérogénéité : les *industry classification misfits*. La littérature précédente a pour habitude d'utiliser les différentes classifications comme des substituts l'une de l'autre, considérant qu'elles groupent les firmes sur la même dimension d'homogénéité. Ici, je prends une approche différente et considère ces classifications comme des compléments l'une de l'autre, en argumentant qu'elles possèdent le même objectif (former des groupes homogènes de firmes), mais qu'elles le font sur des dimensions différentes de l'homogénéité. Ainsi, en étudiant leur convergence j'identifie les *industry classification misfits* par opposition à celles appartenant au cœur de l'industrie (*industry core firms).* Ultimement, je montre les biais qu'engendre l'inclusion des *industry classification misfits* dans les groupes de comparables pour l'estimation des modèles d'*accruals* et la prédiction des *misstatements*.

Dans un second temps, l'objectif est d'intégrer l'utilisation des ratios comptables et financiers pour identifier les firmes hétérogènes. Je pars de la classification qui offre la plus grande homogénéité pour développer une mesure continue d'homogénéité intra-industrie. J'utilise les ratios comptables et financiers qui sont régulièrement utilisés pour mesurer l'homogénéité d'un groupe de firmes. Contrairement aux études précédentes qui utilisent individuellement ces ratios, je propose une approche multidimensionnelle à l'homogénéité. Dans une première

étape, je définis les ratios pertinents pour définir chacune des industries, puis j'utilise simultanément ces ratios pour construire ma mesure continue de distance intra-industrie entre chacune des firmes. Ainsi, je présente les firmes étant les plus éloignées du cœur de l'industrie comme des firmes différenciées (*differentiated firms*). Ensuite, j'étudie les conséquences sur les marchés financiers pour ces firmes. Je montre que les nouvelles d'industries sont incorporées dans les prix des firmes différenciées avec un retard. Aussi, je montre que les analystes couvrent moins ces firmes et commettent plus d'erreurs dans la prédiction des bénéfices de ces firmes. Enfin, je montre que les firmes différenciées souffrent d'une asymétrie de l'information plus importante sur les marchés, ce qui se matérialise par un plus grand écart bid-ask et une action moins liquide.

Enfin, dans un troisième temps, l'objectif est d'utiliser les liens entre les industries pour mieux caractériser les firmes multisegments. Je m'intéresse à une source naturelle d'hétérogénéité intra-industrie — les conglomérats. Par définition, ces firmes opèrent dans plusieurs industries différentes, mais la construction des classifications d'industries restreint leur classification à une industrie. Ceci crée donc naturellement de l'hétérogénéité au sein des industries ce qui amène à les considérer comme complexes, notamment pour les analystes qui se spécialisent par industries. Habituellement, les études précédentes ont considéré que plus une compagnie possède de segments d'affaires différents, plus elle sera complexe. Dans ce chapitre, j'apporte une nuance sur leur complexité en prenant en compte le lien entre les différentes industries dans lesquelles opèrent les conglomérats. Je développe une mesure de distance entre les industries basée sur les ratios financiers. Ainsi, je considère les segments d'affaires comme complexes uniquement ceux qui sont éloignés du cœur d'activité de la firme. Par conséquent, deux conglomérats possédant le même nombre de segments d'affaires peuvent être complexes ou non, dépendamment si leurs activités secondaires sont dans une industrie proche de leur activité première. Ultimement, je montre les conséquences de ces firmes pour les analystes. Mes résultats dévoilent que les analystes ont plus de mal à prédire les bénéfices des conglomérats complexes.

Abstract

This thesis focuses on the selection of peer firms in the context of financial accounting. In this context, the analysis of firms is done cross-sectionally, in comparison with other similar firms – peer firms. Thus, it is necessary to form homogeneous groups of firms for these purposes. Industry classifications represent the most used method because it proposes an objective way to group firms based on their business activities. In addition, they present the advantage of being publicly available and easy to implement. In this thesis, the main objective is to identify sources of intra-industry heterogeneity, and to examine their consequences for several stakeholders. I provide three ways to fulfill this objective.

First, I aim to provide a more complete exploitation of the information provided by industry classifications. Thus, I exclusively use them to identify a source of heterogeneity: *industry classification misfits*. Previous literature tends to consider industry classifications as substitutes for each other, assuming that they group firms on the same dimension of homogeneity. Here, I take a different approach and consider these classifications as complements arguing that they have the same objective (to form homogeneous groups of firms), but that they do it on different dimension of homogeneity. Thus, by studying their convergence I identify firms that are not systematically classified into the same peer group by industry classifications. I refer to them as *industry classification misfits* as opposed to those belonging to the heart of industry (*industry core firms*). Ultimately, I show the consequences of the inclusion of industry classification misfits in peer groups for the estimation of accrual models and the prediction of misstatements.

Then, the main objective is to build on fundamental ratios to identify heterogeneous firms. I start from the classification which offers the greatest homogeneity (GICS) to develop a continuous measure of intra-industry homogeneity. I use accounting and financial ratios which are regularly utilized to measure the homogeneity of peer groups. Unlike previous studies which bring these ratios individually, I propose a multidimensional approach to homogeneity. In a first step, I select the relevant ratios that define each industry. These ratios are then used simultaneously to build my continuous measure of intra-industry distance between each firm belonging to the same industry. Ultimately, I present the firms that are

furthest from the industry core as *differentiated firms*. Then, I study the consequences on financial markets for these firms. I show that industry news is incorporated into differentiated firms stock prices with a delay. Also, I show that analysts are less willing to cover these firms and make more mistakes in forecasting differentiated firms' earnings. Finally, I show that differentiated firms suffer from asymmetric information on the stock market, which occurs as a larger bid-ask spreads and less liquid stocks.

Finally, I aim to account for the industry relatedness to better characterize multi-segment firms. I focus on a natural source of intra-industry heterogeneity - conglomerates. These firms operate in several different industries through secondary business segments, but the construction of industry classifications restricts their classification to solely one industry. Therefore, it naturally creates heterogeneity within industries which leads to consider them as complex, especially for analysts who specialize in industries. Usually, previous studies have considered that the more business segments a company has, the more complex it will be. In this chapter, I add a nuance to this proxy for complexity by considering the relatedness between the industry membership of the secondary business segments in which conglomerates operate. I develop an inter-industry distance based on financial ratios to consider the relationship between industries. Thus, I regard business segments as complex only those that are unrelated to the conglomerate primary business segment. Therefore, two conglomerates sharing the same number of business segments are not systematically equally complex as it depends on whether their secondary activities are in an industry close to their primary activity. Ultimately, I show the consequences of complex business segments for financial analysts. My results show that conglomerates with complex business segments have harder earnings to predict.

Table des matières

Liste des figures, tableaux, illustrations

Liste des abréviations, sigles, acronymes

Abbreviations[1]	Meanings in this thesis
CRSP	The Center for Research in Security Prices
FASB	Financial Accounting Standards Board
FF	Fama-French
FTSE	Financial Times Stock Exchange
GICS	Global Industry Classification Scheme
IFRS	International Financial Reporting Standards
MSCI	Morgan Stanley Capital International
NAICS	North American Industry Classification System
NYSE	New York Stock Exchange
OMB	Office of Management and Budget
S&P	Standard and Poors
SEC	Securities and Exchange Commission
SFAS	Statement of Financial Accounting Standards
SIC	Standard Industrial Classification

[1] In alphabetical order

Remerciements

Ceux qui me connaissent savent à quel point ma pudeur sentimentale est un trait de caractère que je développe quotidiennement, au grand dam des êtres qui me sont chers. Comme vous le remarquerez dans les paragraphes suivants, ce n'est pas (toujours) de la mauvaise volonté. Bien souvent, si je n'exprime pas mes sentiments, c'est simplement parce que je ne suis pas doué pour le faire. Cependant, il me semble que cette thèse marque la fin d'un chapitre important de ma vie (une scolarité de 30 années, ça se célèbre!). Il me parait quand même indispensable de remercier les personnes qui ont contribué à rendre ce doctorat plus agréable. En voici la liste non exhaustive :

- Carl : je n'aurais pu rêver meilleure relation avec mon directeur de thèse. Tu as toujours su trouver les bonnes réponses à mes questions, me soutenir dans les moments compliqués et trouver les bons mots lorsque ma motivation était en berne. Tu as été indéniablement la clé de la réussite dans ce doctorat.

- Daniel : j'ai découvert la recherche grâce à toi, pendant l'été 2013. En quelque sorte, tu m'as donné le virus de la recherche à ce moment-là, et je t'en serais indéfiniment reconnaissant. Tout au long de ma maitrise et mon doctorat, tu auras été un soutien (moral, académique et financier) clé.

- Jean, Alain : vous avoir sur mon comité de thèse est un immense privilège. Votre rigueur (et dureté parfois!) m'a poussé à me dépasser chaque jour. Mon travail a indéniablement bénéficié de votre bienveillance.

- Maryse : tu auras été en quelque sorte ma « grande sœur académique » tout au long de ce doctorat. Ta rigueur (académique!) m'impressionnera toujours. Pour toujours je me souviendrais de nos échanges, nos questions existentielles, et surtout nos « chialages » communs. Je ne te remercierai jamais assez pour tous les conseils que tu m'as pu me donner tout au long de ce doctorat : sans toi, je n'aurais jamais visité tous ces parcs naturels, campings et autres randonnées. Si un jour tu cherches une (re-)reconversion, je saurais déjà où te trouver!

- À mes collègues et amis de la faculté : Janie, Louis-Philippe, Jérôme, Éliane, Michelle, Marion, Maude, Romain, Isabelle, Stéphanie Julie, Myriam ; à travers les bières partagées, les discussions de couloirs ou de bureau, tous les conseils et discussions informelles qui ont rythmé mes journées à la FSA ont permis mon épanouissement quotidien ; continuez à être les bonnes personnes que vous êtes!

- À famille : mes parents, Océane, Jérôme, Orlane, Tom, Léo, Johanna ; votre soutien m'a transmis une force immense. Le soutien indéfectible accordé tout au long de mes (longues) années d'études aura été la clé de ma réussite. La confiance que vous m'avez toujours accordée m'a permis de traverser les épreuves et embuches qui se sont dressées sur mon chemin. Probablement sans le savoir, vous avez été ma source d'inspiration. L'envie de réussir et de vous rendre fiers de moi m'aura poussé à ne jamais abandonner, même dans les moments les plus compliqués.

Introduction

Financial accounting is a discipline of accounting that relies in part on the existence of accounting norms (standards) issued by standard setters. These “norms” (articulated rules) often arise from accounting practices (implied rules) and exist in order to harmonize them (Zarzeski 1996). In other words, they exist to create a uniform framework through which companies report their financial activities, in order to allow users to reduce the cost of processing financial information. However, this harmonization process is risky because it can also make firms look more (economically) similar to each other than they truly are. By extension, groups of firms could then be perceived as *homogeneous* while users of financial statements may prefer financial statements that properly highlight sources of *heterogeneity*, because those may be more useful in providing firm-specific information. Therefore, when issuing standards, regulatory bodies must always deal with the trade off between giving more space to allow financial statement preparers to provide discretionary – *heterogenous* – information and constraining this space to obtain uniform – *homogeneous* – information. This process is complicated by the possibility that stakeholders may have different preferences regarding financial information (e.g. Demski 1973).

In this context, users generally interpret a firm’s financial information by comparing it with a group of firms whose economic activities are assumed to be similar (i.e. *peer firms* or *comparable firms*). To ensure objectivity in peer selection, users typically rely on existing industry classification schemes, in which a group is considered homogeneous if all of its firms are in the same industry, and heterogenous if they are not. The whole objective of my thesis is to study the use of industry classifications in financial accounting. I aim to identify sources of industry heterogeneity and to study its consequences for a multiplicity of stakeholders (investors, financial analysts, or researchers).

The use of industry classifications to select peers has a long tradition in financial accounting. Three main classifications are used by both practitioners and researchers: the Standard Industrial Classification (SIC), the North American Industry Classification System (NAICS) and the Global Industry Classification Standard (GICS). The SIC was first issued in 1937 by the Central Statistical Board of the United States and remains used by the Securities and Exchange Commission (SEC) in 2020. The NAICS was initially issued in the US to replace the SIC in 1997 in every governmental body[2], while the GICS has been issued by a private organization (MSCI and S&P) in 1999. These three classifications are based on a hierarchical structure from broader industries at the higher level to narrow sectors at the bottom. For example, the GICS classification structure uses four levels: the broadest (narrowest) level is referred to as GICS2 (GICS8), where 2 (8) represents the number of numerical digits used to index industries at that level (see Appendix B for an excerpt of the GICS hierarchical structure within two GICS2 industries). Industry classifications have the advantage to offer an objective way to classify firms into homogeneous peer groups. Moreover, they are commonly available in all major public databases (Bloomberg, Compustat, CRSP, etc.) which make their use relatively easy. More precisely, for researchers it enhances the generalization and the replicability of their research. Finally, despite recent challenges from researchers (e.g. Hoberg and Phillips, 2016 ; Lee et al., 2015), the frequent issuance of new classifications by private organizations (e.g. the Industry Classification Benchmark by the FTSE, the Dow Jones Industry Classification System) shows how industry classifications still occupy a central role in the organization of financial markets. The main criticism toward industry classifications is the lack of flexibility toward the constitution of peer groups (Ecker et al. 2013; Hoberg and Phillips 2016). It could result in the creation of *heterogeneous* peer groups which is the opposite of their objective. In my thesis, I aim to investigate how we can adapt and use industry classifications to achieve a better intra-industry homogeneity. Moreover, evidence regarding the consequences of peer groups heterogeneity is relatively scarce. Consequently, the second objective of my thesis is to investigate the implications of heterogeneous firms for financial statement users.

Homogeneity plays a central role in my thesis, even though it is a concept that is hard to distinguish from comparability, which is fundamental to the financial accounting conceptual framework (FASB 1980). Firms reporting under the same set of accounting norms should have comparable financial statements, where comparability is different from uniformity (2018 IFRS Conceptual framework, 2.27). Consequently, two firms experiencing two different economic events should possess comparable financial reports, but not similar ones, enabling financial statements users to compare their disclosures (i.e. earnings, cash flows, accruals, etc.) to make predictions. Since we

[2] The NAICS project was also initiated to uniformize industry classifications with other countries in North America (Canada and Mexico). In the end, the NAICS did replace the SIC in most of the US governmental bodies, except for the SEC.

typically use cross-section comparisons in fundamental analysis to evaluate firms' performance, comparability is a highly desired attribute to financial statements in every context.

I define homogeneity as a broader concept that includes comparability. I define homogeneity as a multidimensional and contextual construct that aims to form groups of firms that can be analyzed simultaneously. I posit that homogeneity should be seen as a multidimensional and continuous construct. I argue that economic forces (e.g. strategy, industry competition) pushes firms to become *heterogeneous* to their peers. I posit that when solely using one industry classification as a peer selection method, we miss the point and do not use all the information available to form the *best* homogeneous group of firms, a significant issue given the known empirical biases of inferences drawn from heterogeneous data (e.g. Owens, Wu and Zimmerman 2017). In other words, we typically create groups of firms containing both *comparable firms* and *heterogeneous firms.* However, the identification of *heterogeneous firms* may be challenging because firms may or may not explicitly disclose these sources of heterogeneity in financial statements. Consequently, the first objective of my thesis is to empirically identify heterogeneous firms.

I posit that industry classifications and financial statements form publicly available information that could be used to identify heterogeneity and avoid mixing up *comparable firms* and *heterogeneous firms* within a given industry. Despite being subject to many horse races, it is still an open debate to determine which classification is *better* suited for accounting (Bhojraj et al. 2003; Hrazdil and Scott 2013) or whether industry classifications are relevant to form peer groups (Lee et al. 2015; Hoberg and Phillips 2016; Ding et al. 2019). Ultimately, I provide a multidimensional view of intra-industry heterogeneity that relies on sources of heterogeneity at both the classification and the firm level. To do so, I develop three research designs that enable to identify different types of heterogeneous firms: *industry classification misfits*, *differentiated firms*, and *complex conglomerates*.

First, by their construction, industry classifications impose the transitivity between firms for peer selection (Hoberg and Phillips 2016). For example, one could argue that firm A might be a competitor of firm B, while firm C is a competitor of firm B but not of firm A. Using industry classifications does not enable this and imposes transitivity. Firms A, B and C will be automatically classified as competitors to each other. Thus, industry classifications impose boundaries to categorize firms in solely one industry which can potentially increase heterogeneity. Moreover, the construction of classifications imposes an equally weighted relationship between firms– firm A is *as much* a competitor for firm B than for firm C. Thus, using industry classifications to select peers provides a binary measure of homogeneity – firm A is (or not) a peer of firm X – rather than a continuous one.

Few studies recognize the existence of intra-industry heterogeneity of industry classifications and its consequences. However, they only expose theoretically the effect of intra-industry heterogeneity (Lee et al. 2015; Hoberg and Phillips 2016) or limit their empirics to a specific context (Peterson et al. 2015; Owens et al. 2017; Ding et al. 2019). In my thesis, I aim to provide more evidence regarding the implications of intra-industry heterogeneity for financial markets. Then, for each type of heterogeneous firms, I choose three different settings to test their implications for both researchers and practitioners. For each source of heterogeneity, I choose a context where the relationship between intra-industry heterogeneity and its consequences is the most straightforward[3].In this thesis, I review prior literature on peer selection methods and intra-industry homogeneity (chapter 1), and then I create three new methods to identify heterogenous firms and illustrate their effects in various settings that are common in accounting research (chapters 2 to 4).

In chapter 2, I investigate how multiple industry classifications can be treated as complements rather than substitutes[4]. I posit that information is lost when a single classification scheme is used to identify peers because each scheme is based on different business dimensions. For example, GICS is based on a firm's ultimate markets (e.g. consumer discretionary), while SIC is based on the production technology (e.g. manufacturing). I use the frequency of listed firms in every possible GICS-SIC combination to distinguish between clusters of *industry core firms* and firms in more unusual combinations that I call *industry classification misfits*.

[3] The main idea behind this choice is to have a setting where the results are less likely to be contaminated by correlated phenomena and to try to rule out as many alternative hypotheses as possible for my results. This is particularly important as two of my chapters (chapter 3 and 4) create proxies based on fundamental ratios which makes it risky to apply them in some contexts (e.g. accrual models).

[4] This is typically what is done when trying to identify which one *performs* better (see for example Bhojraj et al., 2003 ; Hrazdil and Scott, 2013 ; Krishnan and Press, 2003).

First, I test the implications of *industry classification misfits* for accruals model and misstatements predictions. *Industry classification misfits* represent firms that share uncertainty regarding their industry membership. Industry peer groups are required for accrual models in order to provide homogeneous groups of firms sharing the same accruals generating patterns. The objective of accrual models is to identify firms that deviates cross-sectionally from their peers to determine whether their earnings are managed. I use this context because the relationship between intra-industry heterogeneity – as proxied by *industry classification misfits* – and accrual models is direct. Intra-industry homogeneity is an implicit assumption of accrual models, and in this context, I argue that a high uncertainty regarding the industry membership of firms should result in a high uncertainty regarding their (supposedly) shared accruals generating patterns. Thus, *industry classification misfits* have a high probability of possessing a different accrual generating pattern from their industry peers. Ultimately, the outcome of accrual models (i.e. abnormal accruals) is supposedly highly correlated with misstatements. Thus, I test the implications of industry classification misfits for the estimation of accrual models and the use of their outcome to predict misstatements.

My results show that industry classification misfits have significantly larger absolute abnormal accruals than industry core firms, after controlling for an array of firm characteristics such as size, the book-to-market ratio and the volatility of operating cash flows. This result is consistent with the argument that industry-based accrual prediction models offer a poor fit for industry misfits, and the resulting measurement error yields inflated estimates of absolute abnormal accruals. In addition, I show that abnormal accrual estimates for misfits are less contaminated by shocks experienced by other firms in the peer group (Owens et al. 2017) than estimates for core firms. I also find that absolute abnormal accruals are positively associated with future restatements for core firms but not for misfit firms, which suggests that the measurement error for misfits impairs the usefulness of absolute abnormal accruals as a proxy for financial reporting quality. Finally, I show that misfits' stock prices are less affected by industry-wide news than core firms, which indicates that market participants seem to identify misfits to some extent.

In chapter 3, I exploit a source of heterogeneity at the firm-level arguing that firms have incentives to *voluntarily* become heterogeneous to their industry peers in competitive markets. This is usually achieved to obtain a competitive advantage over their competitors and to survive through years. In addition, industry competition can drive firms out of the market and make them appear dissimilar from their peer group, against their will. In this chapter, I develop a continuous measure of intra-industry heterogeneity to identify *differentiated firms* (i.e. firms that are heterogeneous to their industry peers[5]). This measure stems from a three-step methodology. For each industry-year I draw a multidimensional spatial representation of all the firms in the industry where the dimensions are industry-specific fundamental ratios. First, I empirically determine the dimensions that are relevant for each GICS6 industry-year. This step will provide a list of fundamental ratios that characterize each industry-year. I refer to these ratios as "*industry characteristics*". Then, for each industry-year I set a centroid that represents the *prototype* firm of the industry. In the spatial representation of the industry, the prototype firm has the means of the industry-specific ratios (the "industry characteristics") as coordinates. Thus, each industry-year has its own number and set of dimensions to characterize the intra-industry heterogeneity. In the third and final step, I calculate the Euclidean distance between the prototype firm and each firm from the industry. I argue that firms farther (closer) from the industry centroid are *differentiated firms* (industry core firms).

In this chapter, I test the implications of differentiation for financial statement users. I argue that differentiated firms should experience more information processing costs due to their lack of benchmarks. First, I examine how differentiated firms incorporate industry news. I argue that differentiated firms represent a type of complicated firms where the industry component of news is more difficult to incorporate in their prices than for core industry firms (Cohen and Lou 2012). Using daily returns, empirical results show that the association between industry news and firm-specific stock returns is smaller for differentiated firms than for industry core firms. In addition, compared to core industry firms, the association between returns and contemporaneous industry news is lower for differentiated firms, but the association between returns and lagged industry news is higher for differentiated firms. I interpret these results as evidence that investors incorporate industry news in a less timely manner for differentiated firms, due to higher information processing costs. Then, I focus on financial analysts. Previous literature highlights the importance of industry specialization for analysts in order to benefit from economies of scale (Piotroski and Roulstone 2004; Kadan et al. 2012). Since differentiated firms represent firms that

[5] Contrary to chapter 2 where the heterogeneity could come from a *misclassification* of the misfit firms, here I take a different approach. I argue that *differentiated* firms may be very similar to their industry peers in terms of products (or type of business they pursue more generally) despite being heterogeneous to their industry.

are heterogeneous to their industry peers, I argue that providing forecasts for these firms will be costlier for analysts. Thus, I hypothesize that differentiated firms receive less coverage from analysts. Moreover, I predict that forecasting these firms will be harder due to their lack of benchmarks, resulting in less accurate and more disperse forecasts. Results from empirical tests support these hypotheses. Finally, I show how differentiation increases the information asymmetry on the stock market. Firms benefit from their peer information environment in order to help investors to process information disclosed (Peterson et al. 2015; Shroff et al. 2017). Heterogeneity arising within peer groups leads to a greater information asymmetry. Using two different measures of information asymmetry – bid-ask spread and illiquidity (Amihud 2002) – I provide empirical evidence that differentiated firms experience a greater information asymmetry. Taken together, my results highlight the potentially adverse effects of differentiation on capital market participants.

Finally, in chapter 4, I focus on a natural source of intra-industry heterogeneity: conglomerates. Conglomerates are firms owning several business units from different industries. However, through industry classifications these firms are classified in only one industry according to their core business (i.e. the most important business units in terms of sales). Thus, one could argue that all the other secondary business units of the conglomerates represent a source of heterogeneity. Many studies build on this argument and use the number of business units as a natural proxy to measure the operational complexity of multi-segment firms (Hoitash and Hoitash 2018). In this context, operational complexity represents a specific form of heterogeneity. In other words, they consider that, no matter the industry membership of the secondary business units, they always represent a source of heterogeneity. In this chapter, I take a different approach and do not consider all additional segments as heterogeneous. I develop a measure of industry relatedness to assess the incremental complexity of conglomerates' secondary business units. I differentiate between business segments that add complexity to the conglomerates (i.e. heterogeneity to the peer group) and the ones that are more closely related to their core business (i.e. more homogeneous to their peer group). Overall, in this chapter I study the homogeneity both at the firm-level (i.e. whether conglomerates are homogenous to single-segment firms), and at the industry-level (i.e. how industries are homogeneous to each other).

Then, I focus on a particular group of financial statement users (i.e. financial analysts) to test my measure of complexity. Financial analysts have the reputation of being sophisticated participants that have the technical skills to understand complex information. Also, previous literature highlights that analysts specialize in industries in order to issue better forecasts (Boni and Womack, 2006 ; Bradley et al., 2017 ; Clement, 1999 ; Kadan et al., 2012). Since complex firms represent companies having business units less related to their primary business in which the analyst is a specialist, I argue that analysts will face higher information processing costs for these firms. Therefore, I hypothesize that complex firms will experience higher (lower) analyst forecast dispersion (accuracy). Results confirm that analysts' forecasts are more dispersed for complex conglomerates. I provide an analysis regarding the analyst coverage of complex conglomerates. Results confirm that analysts do account for the number of business segments since they are less willing to cover multiple-segment firms. However, I show that analysts fail to recognize when multiple-segment firms are composed of *complex* business segments. Finally, I document a negative association between analysts' forecast accuracy and complexity, suggesting that complex conglomerates earnings are more difficult to predict for financial analysts. However, the effect disappears when the number of business segments is added as a control variable.

The rest of this thesis is organized as follows. First, in chapter 1 I present a literature review on peer-selection methods and intra-industry homogeneity. Then, in chapter 2, 3 and 4 I present the results regarding *industry classification misfits*, *differentiated firms* and *complex conglomerates* respectively. Finally, I draw some concluding remarks in the closing section.

1. Literature review on peer selection methods and intra-industry homogeneity

1.1. Industry classifications

1.1.1. Definition, structure and preliminary evidence

Initially, the first studies to investigate the importance of industries were trying to distinguish between the idiosyncratic and industry components of firms' profitability (Schmalensee 1985). In his study, Schmalensee proposes a simple regression model where the dependent variable is the profitability (return on assets) of a company. The independent variables tested are companies and industries fixed effects and a control for firms' market share in its industry. By comparing the different R² regressions and F-Tests, the author shows that what matters most are the industry fixed effects. This allows him to claim that "industry effects exist and are important, accounting for at least 75 percent of the variance of industry rates of return on assets". Thus, these fixed effects clearly dominate firm fixed effects and firms' market share, indicating that industries have a great explanatory power for the profitability of a business. Additionally, the author was able to show that there are strong inter-industry differences in the profitability of firms. Ultimately, he shows that grouping firms by industry seems a good way to obtain homogeneous groups in terms of profitability.

Until the end of the 1990's, the SIC is dominating the market of industry classifications. It is the only widely available classification despite evidence that it does not seem to provide homogeneous peer groups (Clarke 1989). In his study, Clarke investigates whether the SIC can be used to delineate "economic markets" using three economic variables: *sales change, profit rates* and *stock price changes.* His results highlight that the 1 or 2-digits SIC is useful to create homogeneous groups of firms. However, the 3 and 4-digits SIC - which is narrower - does not seem to increase the homogeneity of the peer groups, compared to the higher level of this classification (SIC1 or SIC2). Therefore, the author concludes that "the SIC does poorly at delineating economics markets". This result deserves to be nuanced since the study only covers the SIC 1000 to 3999 industries which limit its generalization to other industries. Also, in empirical studies (e.g. accrual models) the 2-digit SIC is often chosen since it presents an adequate number of firms to alleviate concerns regarding the lack of degrees of freedom. This study shows that the SIC seems capable of forming homogeneous groups of firms at this level of the classification. However, it does not question the use of the SIC2, but it shows that the SIC as a whole is subject to improvement. Another problem directly related to the SIC is that databases reclassify each firm on criteria specific to their classification methodology. So, a firm can be classified in two different SIC industries according to which database is selected. For example, 38% of firms are classified differently depending on the use of CRSP or Compustat (Guenther and Rosman 1994). It might be an issue from a research standpoint since researchers usually do not disclose the source of the SIC code used (Kahle and Walkling 1996). Kahle and Walking explain that this difference exists because Compustat did not provide historical SIC codes unlike CRSP[6]. However, they show that there is still a large disagreement between the two databases even when using historical codes for both. Using a methodology based on simulations, they show that the differences between the two databases can impact on the results. Ultimately, they show the superiority of the Compustat database over the CRSP for the data on SIC codes.

New evidence adds to the criticism toward the use of the SIC and points out that this classification does not adapt to the creation of new product markets (Hoberg and Phillips 2016). Thus, direct alternatives to the SIC have been issued. These emanated both from organizations currently proposing the SIC classification (or advocating its use), from private organizations, or from academic research. So, to consider the evolution of the economy and changes in the structure of industries, in 1997, The US Office of Management and Budget (OMB) created the NAICS intended to replace the SIC in all government bodies. Like the SIC, the NAICS rely on firms' products and supply chains to classify them. This could be problematic, since two firms sharing the same products and supply chains might not be homogeneous from an accounting perspective. Therefore, the GICS was created by Morgan Stanley Capital International (MSCI) and Standards & Poors (S&P). This new classification takes a financial approach to form homogeneous groups that are based on market perceptions (analysts, investors) and firms' sources of earnings. Thus, it is more in line with financial logic rather than an industrial economics logic (based on products and outputs). Unsurprisingly, the GICS is widely used by researchers in the financial analysts' literature or more generally in finance. Studies have shown that this classification is representative of analysts' choices for the selection of peers (Boni and Womack 2006; Kadan et al. 2012). More precisely, according to these authors, most companies followed by an analyst belongs to the same

[6] Compustat later added historical SIC codes to the database for most firms, beginning with the 1987 fiscal year.

GICS6 industry. This confirms that the GICS seems to correspond well to financial analysts' realities. Conversely, analysts' peer selection appears to be uncorrelated with the SIC industry membership of firms (Ramnath 2002).

Finally, asset pricing research focused on the impact of industries for the Fama-French 3-factor model (Fama and French 1997). The authors find industry cross-sectional variations in the risk factors. Criticism of the SIC classification is implicit in their study, since they only indicate presenting a new way of classifying firms for: "having a manageable number of distinct industries that cover all NYSE, AMEX and NASDAQ stocks" (p. 156). So, these new industries are a reformulation of the classification where 4-digit SIC codes are reorganized to create new industries.[7]

In the same way that studies analyze the relevance of the SIC (Kahle and Walkling 1996; Clarke 1989; Guenther and Rosman 1994) researchers have been interested in these new classifications in order to assess their differences and establish a hierarchy. For example, the NAICS seems very useful for accounting research (Krishnan and Press 2003). Krishnan and Press show that differences between the SIC2 and NAICS3 – the level of these classifications commonly used – are very minimal since only the industries "Services" and "Public administration" represent new industries in the NAICS. This confirms that the NAICS has been introduced to provide an update of the SIC more representative of the changing economy. In their study they use accounting ratios to assess the ability of this classification to create homogeneous groups of firms. They analyze four ratios: profitability (return on assets), liquidity (current ratio), solvency (long-term debt / assets) and asset turnover. The variance of these ratios is estimated for each industry of both the NAICS and SIC. Finally, they compare the mean variance of each classification in order to determine which one achieves the best intra-industry homogeneity. Their analysis shows the ability of the NAICS to increase the intra-industry homogeneity for certain industries (manufacturing, transportation, and services). Also, the NAICS has a real impact on the lower level of classification – levels comparable to the SIC3 or SIC4 – which was a concerned before (Clarke 1989). Finally,

[7] The most commonly used classification contains 49 industries, usually referred as FF49. Other narrower and broader reclassifications are available on Kenneth French's website (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html). However, they are less exploited because they classify a significant portion of firms into the "uncategorized" industry group. This limits their usefulness.

the authors replicate a previous study. They show that the choice of a classification over another can have an impact on research results.

The emergence of the NAICS and GICS opens up new possibilities for studies using industries to form homogeneous groups of firms. Despite sharing the same objective, they do not seem to be equivalent and some give better results than others where the GICS seems to be the most homogeneous one and is favored in capital market research (Bhojraj et al. 2003; hereafter BLO). In their study, BLO are the first to be interested in industry classifications beyond the SIC and NAICS, integrating the GICS classification and the classification from Fama and French (FF). They are making contributions to both finance and accounting literature since they are evaluating the impact of the choice of a classification for asset pricing and financial analysis. In the same way as Krishnan and Press (2003) are interested in the convergences between the SIC and NAICS, BLO study the concordance between the classifications in a broader way since they take into account four classifications (SIC, GICS, NAICS, FF). In their results (Table 2), they show that the NAICS and FF are very close to the SIC, while the GICS seems less related. For example, only 34% of the firms belonging to the SIC36 industry "Electronic And Other Electrical Equipment And Components, Except Computer Equipment" are classified into the same GICS industry (GICS 452020 "Technology Hardware, Storage & Peripherals"). This means that according to the GICS, 66% of the firms of the SIC36 industry are not classified with the *right* peer group, resulting in a potential increase in intra-industry heterogeneity. In the end, they show that the degree of correspondence between the SIC and the GICS, NAICS and FF is 56%, 80% and 84%, respectively. For the NAICS and FF, the high degree of correspondence is not surprising since they are built on the same foundations. On the other hand, the result regarding the GICS is more appealing. Even if they rely on different criteria to classify firms, we do not expect such a low degree of correspondence between the classifications. This suggests that despite having the same objective (i.e. form homogeneous peer groups), they do so on different dimensions. Finally, in the rest of their study BLO show that the GICS seems to better explain industry cross-sectional returns, valuation multiples (price-to-book ratio, enterprise value-to-sales, price-to-earnings ratio), financial accounting ratios (return on net operating assets, return on equity, asset turnover, profit margin) and other financial ratios (long-term analyst growth forecasts, one-year ahead realized sales growth). BLO concludes that the GICS is the most homogeneous industry classification and should be used in accounting or finance to select peers. This superiority is mainly explained by the financial orientation taken by this classification to delimit the industries, while the other classifications are more "product" oriented. Nevertheless, through the interpretation of their results, they assume that the classifications are substitutes for each other. They claim that it is possible to prioritize one classification over the others in any context by choosing the "best" industry classification – the one that offers the best intra-industry homogeneity. However, this assumption is debatable given the foundations on which each of the classifications are based. Even if the GICS provides the best intra-industry homogeneity overall, we note that it is not the case for every dimensions of heterogeneity (leverage for example). Thus, using the GICS might not be the best possible choice in every context. In the same way that there can be several dimensions to "good governance" (Larcker et al. 2007) or to "earnings quality" (Dechow et al. 2010), I think that intra-industry heterogeneity can also articulate around a multi-dimensionality of its proxies given that each of the industry classification was built to meet specific needs and relies on different assumptions.

1.1.2. Improvements of intra-industry homogeneity using accounting information

Researchers propose ways to alleviate concerns regarding the intra-industry heterogeneity of industry classifications using accounting information. For example, the operational cost structure of firms can be used to form homogeneous groups of firms for auditors (Cairney and Young 2006). In this context, the authors are trying to identify homogeneous industries from an auditor standpoint. They argue that the most homogeneous industries are the one where firms share the greatest correlation on their operational cost structure. Ultimately, they want to show that auditors benefit from greater economies of scale when auditing multiple firms belonging to the same *homogeneous* industries, resulting in a higher market concentration of audits. They present their peer selection method as a way of forming *economically* homogeneous groups of firms since : "*the rates of change in the operating expenses of homogeneous firms are similar and reflect the underlying similarity of operations because concurrent economic conditions have resulted in a similar reported financial impact on these companies*". In other words, they implicitly claim that their methodology could be used to form peer groups that are *economically* homogeneous (i.e. homogeneous in every context). Using these approach to peer selection could be problematic since it represents only *one* dimension of homogeneity that is subject to limitations. For example, firms could experience changes in their cost structure due to organizational or strategic issues inside the company (Owens et al. 2017). This would make them appear dissimilar to their peers while they may operate on the same market. In addition, this methodology does not seem to suit the *new* economy (i.e. based on business services, technologies) and seems more adapted to the *old* manufacturing economy where the relationship between companies' operations and their cost structure is more straightforward.

Studies looking at firms' earnings response to economic news – as measured by stock returns – has a long tradition in accounting research (Basu 1997 ; Parrino 1997 ; De Franco et al. 2011). Industries that exhibit a high correlation on their firms' stock returns are presented as the most *homogeneous* ones (Parrino 1997). From an accounting standpoint, two firms are comparable if their financial reporting system presents the same response to a given economic news (Basu 1997 ; De Franco et al. 2011 ; hereafter "DKV"). In their study, DKV propose a new measure of intra-industry heterogeneity based on their definition of comparability. They assume that economic news impact earnings in a timely manner which is consistent with Basu (1997) for negative news but it seems very unlikely for *positive* news since the latter usually proxy for earnings surprises or future economic performance. Also, industry news could be integrated differently into firms' financial reporting systems for many reasons. For example, two firms could experience a similar response in their earnings to two independent economic news, which will make them appear comparable even if the two economic news are unrelated. Also, these two events could have different implications for these firms' earnings resulting in an increase in intra-industry homogeneity since these firms will not be considered comparable. More generally, using the firm news – stock returns – as a determinant of firms' earnings in order to form homogeneous group of firms seems risky. It looks hard to isolate clean settings of news where the comparability of firms could be properly assessed. Thus, the relationship between stock markets and financial reporting systems looks too complicated to be directly used to measure the intra-industry homogeneity.

Despite being subject to discussions, the comparability measure proposed by DKV opens lots of avenues for research. While DKV present the output-based (i.e. based on earnings) methodology used as a strength of their study, others regard this as weakness and seek alternatives (Peterson et al. 2015). Peterson et al. (2015) present an input-based proxy for comparability relying on firms' 10-K disclosures. They use textual analysis to measure firms' accounting *consistency* based on firms' *accounting policy disclosure* similarities. More interestingly, they also propose a new measure of comparability based on the textual analysis of firms' business descriptions. For both *consistency* and *comparability,* they provide a cross-sectional and time-series construct. Even if their primary objective is to extend previous work by DKV, they achieve a different goal and provide evidence regarding another dimension of heterogeneity through their comparability measure. While the study of DKV build on the integration of economic news by financial reporting systems, Peterson et al. (2015) contribute to the literature through a focus on the firms' textual disclosures.

Companies' size plays a role in explaining their performances and other outcomes. Most of the empirical studies include companies' size as a control for firm operational complexity at the firm-level (e.g. Dechow and You (2012)). Thus, size appears as a natural criteria to form homogeneous groups of firms (Albuquerque 2009; Ecker et al. 2013). In the context of accruals, the formation of peer groups based solely on firms' size seems to perform better than industry classifications (Ecker et al. 2013). For these authors, size looks to be a better criterion to select peers than industry membership: "*We consider size because, as we explain in more detail later, it is an intuitively grounded alternative (to industry membership) indicators of similarity; that is, a group of larger firms is more alike than is a mixed group of larger and smaller firms*" (p. 191). They give three main arguments for the use of size as a unique dimension to form homogeneous groups. First, firms of similar size should experience the same growth rates where the largest firms are those with a lower growth compared to smaller firms. In addition, they argue that the largest firms are also those having the most complex operations (i.e. they likely have more business units) or are the most monitored (higher coverage from analysts, audited by a big 4, presence of institutional investors). Thus, firms of similar size share common characteristics which lead authors to assume that they should share the same accrual-generating pattern. Using simulations, the authors test the ability of accruals models to detect earnings management across various peer selection methods. Peer groups based on the previous year's total assets (lagged total assets) are those that provide the highest detection rate of earnings management. Conversely, groups based on total assets for the current year do not provide satisfactory results. This is appealing since the two variables used to form peer groups (i.e. lagged total assets and total assets) theoretically represent a similar economic concept (i.e. companies' size). This means that the relationship between the model estimated and the peer groups formed is more complicated than what is exposed. Overall, it questions whether the use of size as the unique dimension to form economically homogeneous groups of firms is adequate.

Albuquerque (2009) takes a different approach to account for companies' size. In her study, size is used as a secondary criterion to determine peer firms within SIC industries. By adding this variable, the author provides a refinement of intra-industry homogeneity through industry classifications. Hence, it considers the intra-industry heterogeneity existing within the SIC2 and SIC3 industries and provide a way to control for it. The use of the size variable seems more appropriate in this context, as it is not used as a substitute for the use of industry classifications. However, this can create problems in estimating empirical models. The addition of a second criterion decreases the number of observations in peer groups which raise concerns regarding the (lack of) degree of freedoms.

Overall, the use of accounting information to increase the intra-industry homogeneity looks interesting at first sight. However, it could give rise to new potential issues. First, when using the accounting information as a substitute for industry classifications, it may offset the economic meaning of peer groups. For example, in Ecker et al. (2013) the authors show how this measure of homogeneity may be of interest in the context of accruals, but it seems dangerous to give it a broader meaning. Also, even if it provides homogeneous groups that allows a better detection of earnings management, the use of *lagged total assets* as a sole dimension to heterogeneity leads to a reduction in the explanatory power of accruals models, compared to the use of SIC industries. Consequently, the peer selection method proposed contains a limited economic meaning making its usefulness limited outside the context of their study (earnings management detection). Accounting information has generally been used to increase the intra-industry homogeneity in some specific contexts. Though, only *one* dimension of accounting information – one variable (e.g. size, operating expenses) or a broader concept (e.g. comparability) is provided. Recently, a study from Ding et al. (2019) provides a multi-dimensional approach to peer selection using machine-learning. They use a methodology based on k-medians clustering to constitute groups of homogeneous firms using financial ratios. They want to use their peer selection method to be able to better predict bankruptcy and misstatements. Thus, they choose financial ratios that are related to these topics. Again, despite being innovative through their multidimensional approach, their results are highly contextual. Their peer selection method is unlikely to provide homogeneous groups firms outside the context of their study. Thus, it lowers its value and reduce its generalization power, although it opens avenues for future research.

1.2. Direct alternatives to industry classifications

Several studies address the weaknesses and limitations of industry classifications, but few direct alternatives have been proposed. The main criticism toward these classifications remains their lack of evolution over time and their rigid hierarchical structure (Hoberg and Phillips 2010; Hoberg and Phillips 2016). In their study Hoberg and Phillips (2010) propose a new peer selection method based on textual analysis. They use cosine similarity as a technique to analyze the business description from 10-K and to evaluate the product similarity between each pair of firms. They calculate a yearly measure of similarity for each pair of firms which give them a very dynamic peer selection method. Moreover, it provides a continuous measure of intra-industry homogeneity for every firm-year. Also, their structure is less rigid than industry classifications since it enables the creation of peer groups of any size. This could be useful for contexts where models are estimated by industry since sample size could be adjusted easily to the number of degrees of freedom required. Thus, it could be useful to accruals models where the use of industry classifications seem questionable (Ecker et al. 2013). Thus, their classification system is based on a scoring that allows the creation of industries of any size, or any minimum scores (i.e. minimum level of intra-industry homogeneity).

Due to its construction orientation, the GICS seems to be the classification that best represents the choices made by analysts for the selection of peer firms (Boni and Womack 2006; Kadan et al. 2012). Before the introduction of the GICS, Ramnath (2002) seeks to understand the spillover effect of firms' earnings announcements on its industry peers through analysts' forecasts. However, according to him, the SIC is not representative of the analysts' peer choices. Thus, the author proposes a new peer selection method. Within each SIC industries, subgroups of firms sharing at least 5 analysts are formed. This methodology is interesting in the context of this study where information spillovers are examined to test the market efficiency hypothesis. However, restrictions on analysts can be problematic in other contexts. Again, quantitative accounting and financial research require large samples for degrees of freedom needs. However, this methodology leads to the creation of very small groups of firms. In addition, the most covered firms by analysts are also the largest in size. A selection bias on size could appear since small firms will be excluded from the sample since analysts are less willing to follow them. This is even more problematic because of the tension in the use of size as the sole criterion for selecting peers. If we consider that size can be a selection criterion, then the methodology proposed by Ramnath (2002) is problematic because it excludes small firms, when they could simply be included in the same peer group. Conversely, if size is not an exclusive selection criterion, two firms of different sizes can be compared. However, with this methodology, small firms will have little chance of being grouped with larger firms.

Analyst coverage contains information about economic linkage between firms that could be used for peer selection (Ali and Hirshleifer 2020). In their study, Ali and Hirshleifer (2020) use analyst coverage to create a network of firms. They analyze the shared coverage from analysts between two firms as a measure of similarity, which enable them to ultimately form peer groups. Their methodology presents similar advantages that a continuous measure of homogeneity offer identified in other studies (Lee et al. 2015; Hoberg and Phillips 2016). However they suffer from a weakness identified by De Franco et al. (2015). The study by De Franco et al. (2015) shows that analysts may use their discretion to select peers to match their personal incentives. For example, the links maintained with investment banks can force analysts to choose peers that fit their need in terms of evaluation, rather than choosing the most homogeneous firms. From a peer selection perspective, De Franco et al. (2015) identify new limits to the methodology proposed by previous studies using analysts (Ramnath 2002; Ali and Hirshleifer 2020). The authors show that analysts' peer choices can be biased by factors that are not economically linked to the fundamentals of firms but may be motivated by personal incentives.

Data gathering by users on government platforms (EDGAR) contains information on peer selection (Lee et al. 2015). The authors analyze the sequence of document downloads by users. They argue that investors will sequentially download the reports of firms Y and Z considered as peer firms of firm X to carry out their financial analysis of firm X. For each firm, they create a network of peer firms based on the sequence of download of financial reports. Thus, they create a scoring system to measure the distance between one firm and another. The advantages related to this peer selection method remain similar to those proposed by the study of Hoberg and Phillips (2016): dynamic over time; flexibility in the constitution of industries. However, there are some limitations. First, the methodology relies on access to proprietary data to construct the classification (EDGAR Log details), which makes its replication almost impossible. Second, there are many doubts that investors do download 10-K reports, especially directly from government platforms like EDGAR (Loughran and McDonald 2017). More generally, little information is available on who actually use this type of platform. For example, if the score is mainly based on downloads from unsophisticated investors it may result in a noisy measurement. Likewise, behavioral finance has shown the existence of a “familiarity” bias where users tend to pay more attention to stocks they know, or which relate to products they can buy. This bias can have the consequence of directing their research of a peer company, without considering the one offering the best economic homogeneity.

2. Industry classification misfits

2.1. Introduction

A large body of archival research in accounting and finance investigates the association between firm-specific characteristics and various outcomes such as performance or financial reporting quality (FRQ). A classical concern is whether measured associations are driven by actual economic phenomena rather than model specification issues, such as correlated omitted variables.

One way to mitigate this concern is to take into account industry membership. A typical approach is to decompose a variable between expected and unexpected components, in which the resulting "unexpected" (abnormal, discretionary) component measures how the firm deviates from the industry norm at a given point in time. A popular example is the prediction of accruals (e.g. Dechow et al., 1995; Kothari et al., 2005), in which the decomposition is generally based on industry-year regressions.[8] Intra-industry homogeneity is therefore critical for the validity of subsequent inferences and some researchers have illustrated the adverse effects of heterogeneity (e.g., Owens et al., 2017).

In this chapter, I argue that differences between industry classification schemes can be exploited to improve peer firm selection and intra-group homogeneity. More precisely, I argue that the leading classification schemes are *complements* rather than *substitutes*. I posit that since existing schemes are based on different classification criteria and for different users, relevant information is lost when a single classification is used to identify groups of "similar" firms. To illustrate this argument, I combine information from multiple classifications in order to isolate *industry classification misfits* from other firms (i.e., core firms). I then describe how misfits differ from core firms on various dimensions (e.g., abnormal accruals, return comovement) and provide supplementary analysis to ensure that our results are driven by the misfit effect (i.e., heterogeneity) rather than alternative explanations (e.g., quality, risk).

My misfit identification scheme is simple to implement and replicate. The scheme is based on six-digit GICS (hereafter GICS6) and two-digit SIC (SIC2) industry levels. Both are available from

[8] Similar approaches have been used to distinguish industry-level from firm-specific information in stock returns (e.g. Piotroski and Roulstone 2004) or earnings (e.g., Hui, Nelson and Yeung 2016), or to identify firm-specific investments in intangibles (Enache and Srivastava 2018).

Compustat, and Bhojraj et al. (2003) show that these levels result in a comparable number of industries. Every year, I tabulate the number of firms for each GICS6-SIC2 combination. Then within a given GICS6 industry, I define core firms as those with a *reasonably frequent* GICS6-SIC2 combination; firms with other (alternative) combinations are considered misfit firms. The intuition is that for a given consumer market (GICS6), firms with reasonably frequent production processes (SIC2) define the "true" industry core, while firms with alternative production processes are misfits.[9]

2.2. Hypothesis development

2.2.1. Accruals models and intra-industry homogeneity

A pervasive characteristic of archival accounting research is the widespread use of the abnormal accruals construct. In general, abnormal accruals are defined as a deviation from the industry norm and obtained using a prediction model (e.g., Jones 1991; Dechow et al. 1995; Dechow and Dichev 2002) estimated within industry and year. An interpretation is that high abnormal accruals may be signs of earnings management (or, at a minimum, low earnings quality). The *intra-industry homogeneity assumption* is therefore fundamental to the process as heterogeneity generates measurement error in abnormal accruals and potentially flawed inferences in a subsequent stage. However, Owens et al. (2017) show that this assumption is unlikely to hold in most settings, as discrete firm-specific operational shocks generate significantly more imprecise estimates of abnormal accruals and contaminate abnormal accrual estimation for the whole industry in the period during which the shock happens and in the following periods. In addition, Peterson et al. (2015) show that firms whose textual disclosures in financial statement notes are dissimilar to other firms in their industry have higher absolute *abnormal* accruals, a finding which they interpret as evidence of poor model fit rather than lower earnings quality.

[9] I deliberately use the imprecise term "reasonably frequent" to emphasize that one of the most significant design choices required by my method is the determination of "how big" the industry's core should be, and how many misfits should remain. My main design defines a "large" core, where all GICS6-SIC2 combinations that individually account for at least 5% of all firms in a GICS6 industry are considered part of the core. In general, results are qualitatively similar with a "smaller" core (i.e., more misfits); see additional analysis. Results are also qualitatively similar when three-digit NAICS (NAICS3) is added or when the core/misfit distinction is based on a SIC2 "anchor" instead, e.g., reasonably frequent GICS6 -SIC2 combinations for a given SIC2 industry. I report the main results with GICS6 as the "anchor" because Bhojraj et al. (2003) and Hrazdil and Scott (2013) provide evidence of GICS superiority, which makes it more difficult to demonstrate the benefits of using additional industry classifications.

These findings suggest that the identification of bad peers is important because heterogeneity concerns may impair researchers' abilities to draw conclusions based on accruals. For example, it is difficult to interpret whether the absence of a statistical association between abnormal accruals and the probability of having a restatement is due to the absence of a "true" association or noisy estimates of abnormal accruals.[10]

2.2.2. Industry classification misfits

Any classification scheme that assigns all units to a predetermined number of groups will create observations that are more difficult to classify and that become sources of intra-group heterogeneity. I label these observations *industry classification misfits*, and I label other firms *core firms*. As misfits are different from core firms, I argue that they experience a different accrual generating pattern. Because normal accruals are based on an industry-level regression, abnormal accruals (e.g., the residual) are measured with an error for misfit firms, yielding higher *absolute* abnormal accruals. In other words, while Owens et al. (2017) argue that accrual estimation is temporarily biased by firms' operational shocks, I argue that accrual estimation is permanently biased by intra-industry heterogeneity driven by classification misfits. This discussion yields the following hypothesis:

H2.1: Absolute abnormal accruals are higher for industry classification misfits than for industry core firms.

Prior literature has demonstrated that accruals are associated with earnings management or fraud (e.g., Jones et al. 2008), but that industry-based abnormal accrual estimation surprisingly provides no additional benefit over basic accrual components, such as change in receivables, for the prediction of future accounting-related SEC enforcement actions against a reporting entity (Dechow et al. 2011). I argue that estimation error is a contributing factor. For industry classification misfits, the estimation error may even lead to the absence of a significant association between abnormal accruals and misstatements. For core firms, the problem is likely not as severe, despite the possibility of "contamination" of abnormal accruals through the presence of misfits in the peer group (e.g., Owens et al. 2017). To investigate this issue, I will use two proxies of

[10] The converse is also true, as an observed significant statistical association may be the result of correlation between the investigated construct and measurement error in accruals rather than a "true", economic, association.

accounting misstatements – future restatements and SEC *Accounting and Auditing Enforcement Releases* (AAER) – to test the following hypotheses:

H2.2: The association between absolute abnormal accruals and accounting misstatements is lower for industry classification misfits than for industry core firms

Previous studies show the importance of industry information as a component of firms' stock price movements (e.g. Roll 1988; Durnev et al. 2004). Since misfit firms represent firms heterogeneous to their industry peers, I argue that the extent to which industry returns help predict the stock returns is lower for misfit firms than for core firms. Thus, I formulate the following hypothesis:

H2.3: The association between firms' stock returns and industry returns is lower for industry classification misfits than for industry core firms.

2.3. Research design

2.3.1. Construction of *MISFIT*

In this chapter, I question the traditional "all-or-nothing" approach to the use of industry classification schemes, in which all firms in the same (in another) industry are considered equally good (bad) peers. Instead, I posit that homogeneity is a multidimensional construct and that firms similar on many dimensions are better peers than firms that are similar on fewer dimensions. I further argue that the imperfect overlap between different industry classification schemes suggests that each scheme captures different dimensions of homogeneity, and that information from multiple schemes can be combined to distinguish between core firms and misfits. Under this view, the argument that an industry classification is necessarily *better* than others raises the risk that information contained in a "lesser" classification will be discarded even if it is easily accessible and incrementally relevant to the identification of peers. In other words, I posit that industry classifications can be treated as complements rather than substitutes, and I exploit the variation in the GICS and SIC classifications to identify classification misfits.

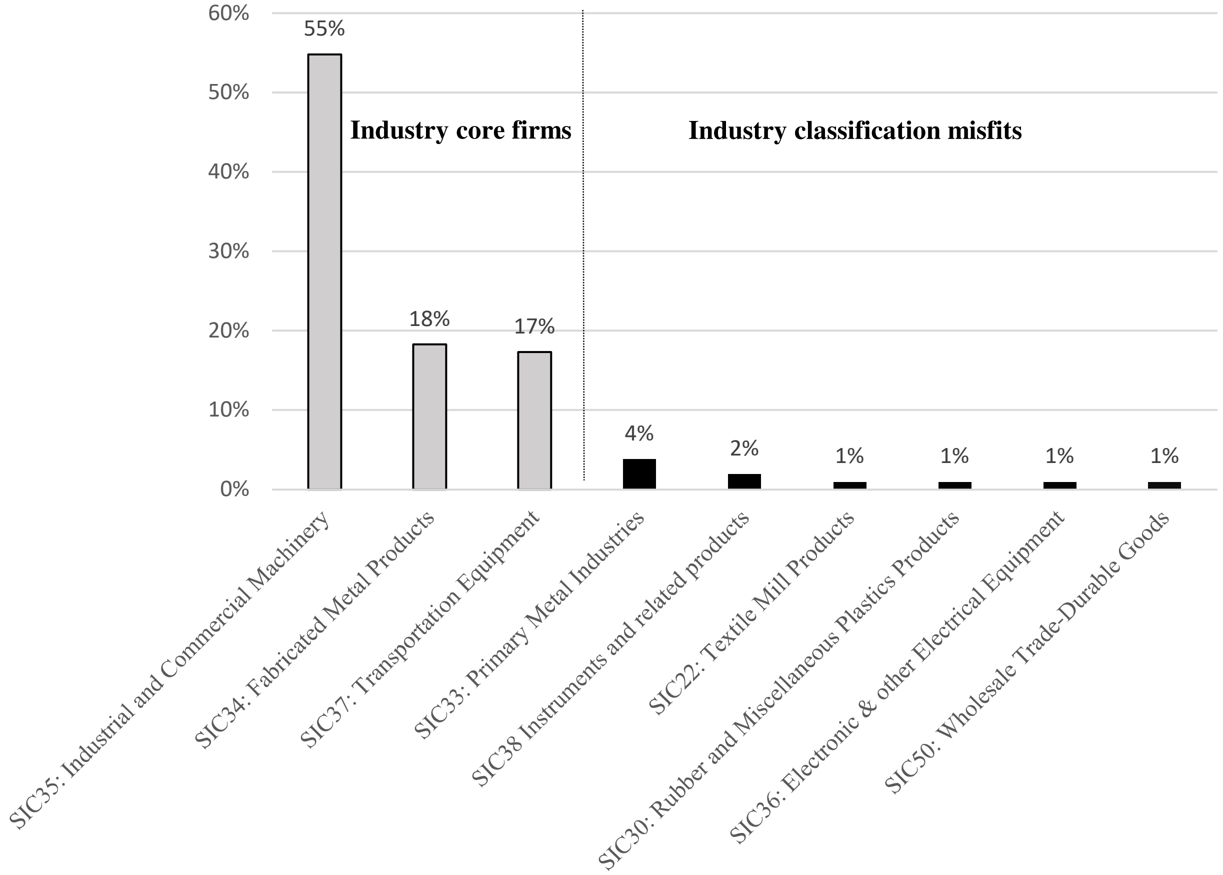
The simplest approach to demonstrate complementarity between industry classification schemes is to interact two schemes; I choose GICS6 and SIC2. I choose GICS6 because it has the least overlap with other schemes (Bhojraj et al. 2003), and also because it is the only "market-oriented" classification in the sense that industry groups are defined according to (customer) markets. I use SIC2 because it is the most widely used of the "production-based" classification schemes.[11]

My methodology to identify misfits and core firms is similar in spirit to the method used by Bhojraj et al. (2003) to measure the degree of convergence between different classification schemes. Compustat assigns GICS6 and SIC2 codes for all observations (firms); all firms therefore have a GICS6-SIC2 combination which can change over time as firms evolve. Each year, I calculate the number of firms for each GICS6-SIC2 pair. I then classify firms in a GICS6-SIC2 pair as *misfits* (*core firms*) if their pair accounts for less than 5% (at least 5%) of all firms in the same GICS6 industry during that year.

Figure 2.1 shows the distribution of firms in the *gics201060* "Machinery" industry across SIC2 industries for the year 2015 (n=106 firms). The most common GICS6-SIC2 combinations that contain *gics201060* involve three SIC2 industries, *sic35* "Industrial and Commercial Machinery", *sic34* "Fabricated Metal Products", and *sic37* "Transportation Equipment", which respectively account for 55%, 18% and 17% of all firms in the *gics201060* industry. Firms with these three combinations are *core firms* (*MISFIT*=0). Ten other firms have a *gics201060*-SIC2 combination that individually accounts for less than 5% of all *gics201060* firms; these ten firms are *misfits* (*MISFIT*=1).

[11] Results are qualitatively similar if NAICS3 or 48 Fama-French industries are used instead of SIC2.

Figure 2.1: Distribution of the industry GICS 201060 “Machinery” across SIC2 industries for the year 2015



Two important choices underlie this research design. First, because the GICS6 classification generates more homogeneous groups than other schemes (Bhojraj et al. 2003; Hrazdil and Scott 2013), I use GICS6 industries as an “anchor” in the sense that core/misfit combinations are determined within a given GICS6 industry. In other words, with my methodology, I investigate whether the SIC2 classification contains valuable information beyond that contained in the presumably superior GICS6.[12] Second, the 5% frequency cutoff I use to classify firms as core or misfit firms implies that all in a *reasonably frequent* GICS6-SIC2 combination are considered core firms. I choose a low cutoff to make sure only firms in the most unusual combinations are considered misfits. For example, in the *gics201060* industry for 2015, I consider firms in the *sic34*

[12] In untabulated tests, I identify misfits and core firms with SIC2 industries as anchors. The results show larger differences between misfits and core firms than those presented in this chapter, consistent with the argument that the SIC2 classification generates more heterogenous groups and that information from another classification scheme (GICS6) can be used to identify the firms that contribute to intra-SIC2 heterogeneity.

and *sic37* industries as core firms, even though they are not in the most frequent *gics201060-sic35* combination.

2.3.2. Empirical models

2.3.2.1. Accruals

In order to test my predictions regarding absolute abnormal accruals, I use the nonlinear model from Ball and Shivakumar (2006; nonlinear i.e., *NL*). I use the *NL* model in main tests because it controls for the effect of current performance on accruals (e.g., Dechow, Kothari and Watts 1998; Kothari, Leone and Wasley 2005) without using a variable directly affected by accruals as a proxy for performance (e.g., return on assets). This model is estimated by year and industry (GICS6):

$$TACC_t = \alpha + \beta_1 CFO_{t-1} + \beta_2 CFO_t + \beta_3 CFO_{t+1} + \beta_4(\Delta SALE_t – \Delta RECT_t) \qquad (2.1)$$

$$+ \beta_5 PPEGT_{t-1} + \beta_6 DCF_t + \beta_7 DCF_t * CFO_t + \varepsilon_t$$

All variables are defined in Appendix A. To test H2.1, I use the absolute value of the residual in this model, $|DACC_{NL}|_t$, as the dependent variable in the following regression model:

$$|DACC_{NL}|_t = \alpha + \beta_1 MISFIT_t + \beta_k [Controls] + \varepsilon_t \qquad (2.2)$$

If abnormal accruals of misfit firms are measured with error in the first-stage industry-based regression, the coefficient on *MISFIT* in the second-stage Eq. (2.2) should be positive, consistent with H1. Control variables in Eq. (2.2) include all regressors from the first-stage model (Chen et al. 2018) and determinants of absolute abnormal accruals according to prior research (e.g., Jones et al. 2008; Peterson et al. 2015).[13] These include return on assets (*ROA*), total assets (*Size*), the standard deviation of cash flow from operations and sales ($\sigma(CFO)$ and $\sigma(Sales)$), book-to-market

[13] I do not directly use total accruals as the dependent variable in a one-stage model because I have no basis on which to expect that misfit firms manipulate earnings upward in a more (or less) significant manner than core firms.

(*BtoM*), and indicators for the presence of a new auditor (*ChAuditor*), recent debt or equity issues (*Issue3y*) or a Big 4 auditor (*Big4*). Controls also include year fixed effects[14].

Owens et al. (2017) show that firm-specific shocks cause measurement error in abnormal accruals, both for firms experiencing the shock and for other firms in the same industry ("other-firm contamination"). I expect *MISFIT* to represent a different source of measurement error because industry membership is permanent while shocks are temporary. Nevertheless, I estimate a model that includes shock-related variables:

$$\begin{aligned}|DACC_{NL}|_t = \quad & \alpha + \beta_1 MISFIT_t + \beta_2 Peer\_shock_{t-1} + \beta_3 Peer\_Shock_{t-1} * MISFIT_t \\ & + \beta_4 Idio\_Shock_{t-1} + \beta_5 Op\_Shock_t + \beta_k [Controls] + \varepsilon_t \end{aligned} \qquad (2.3)$$

Compared to Eq. (2.2), all three new variables in Eq. (2.3) are defined as in Owens et al. (2017). I include two proxies for same-firm shocks (Op_Shock_t and $Idio\_Shock_{t-1}$) to ensure that the association between $MISFIT_t$ and abnormal accruals is not driven by more frequent shocks among misfit firms. As for $Peer\_Shock_t$, a proxy for the prevalence of shocks to other firms in the same GICS6, I want to determine whether the "other-firm contamination" issue affects both industry misfits and core firms. To illustrate this, I interact $Peer\_Shock_t$ with $MISFIT_t$. My intuition is that the other-firm contamination issue is more severe for core firms than misfits, because the latter's abnormal accruals already suffer from poor model fit due to an imperfect industry classification. If this is true the coefficient on $Peer\_Shock_t * MISFIT_t$ will be negative.

2.3.2.2. *Misstatements*

My tests on misstatements focus on the consequences of measurement error in abnormal accruals. If abnormal accruals are measured with error for industry classification misfits, then the association between $|DACC_{NL}|_t$ and proxies for misstatements will be weaker for misfits than core firms. I use future restatements ($Restate_t$) and future accounting-related enforcement actions by the SEC ($AAER_t$) as proxies to indicate misstated financial statements. I use the same model for both proxies, except for the dependent variable; the following model is used with $Restate_t$:

[14] I do not include industry fixed effects in these results since they are highly correlated with *Peer_Shock.* Although, in this model I do not include *Peer_Shock*, the main interest of these results is the comparison between the models (i.e. between Eq 2.2 and Eq 2.3). Thus, for comparison purpose I do not include industry fixed effects in any models. In untabulated results, I include them in Eq 2.2 and show that their inclusion does not change the results.

$$Restate_t = \alpha + \beta_1|DACC_{NL}|_t + \beta_2 MISFIT_t{*}|DACC_{NL}|_t + \beta_k[Controls] + \varepsilon_t \quad (2.4)$$

If absolute abnormal accruals proxy for low accounting quality, they will be positively associated with future restatements. Consistent with H2.2, I expect a negative coefficient on $MISFIT_t{*}|DACC_{NL}|_t$, because measurement error in abnormal accruals impairs the usefulness of accruals for the prediction of misstatements.

2.3.2.3. Return comovement

My test of H2.3 focuses on the association between industry returns and firm stock returns. H2.3 predicts that the incorporation of industry news in stock prices is lower for misfit firms than for core firms. Following previous literature (Cohen and Lou 2012), I estimate the following model (firm *i* subscripts omitted):

$$\begin{aligned} RET_t = \; & \alpha + \beta_1 INDRET_t + \beta_2 MISFIT_t{*}INDRET_t + \beta_3 INDRET_{t-1} \quad (2.5) \\ & + \beta_4 MISFIT_t{*}INDRET_{t-1} + \beta_5 RET_{t-1} + \beta_6 MISFIT_t{*}RET_{t-1} + \beta_7 MISFIT_t \\ & + \beta_k[Controls] + \varepsilon_t \end{aligned}$$

In Eq. (2.5), daily firm returns (RET_t) are regressed on contemporaneous value-weighted GICS6 industry returns ($INDRET_t$), the interaction between industry returns and $MISFIT_t$, and control variables.[15] β_1 captures the proportion of industry-related information in a firm's stock returns and is assumed to be positive. If misfit firms' stock prices contain less industry-related information than core firms' stock prices, β_2 will be negative. Eq. (2.5) includes lagged industry and firm returns (respectively $INDRET_{t-1}$ and RET_{t-1}) to account for the possibility that stock prices react to industry news with a lag and for autocorrelation in daily returns, and interactions between $MISFIT_t$ and these two variables. If investors are unsure about misfit firms' industry membership and need more time to incorporate industry news in stock prices, β_4 will be negative. Other control variables include lagged industry returns ($INDRET_{t-1}$), lagged same-firm return (RET_{t-1}), total assets (*Size*), book-to-market (*BtoM*) and turnover (*Turnover*).

[15] For any given firm *i*, $INDRET_t$ is based on the stock returns of all other firms in the same industry.

2.3.3. Sample construction

In order to complete my analysis, I use data from different databases to create an initial sample common to all chapters. First, I obtain fundamentals data regarding financial statements through Compustat for both firm-level and segment-level data. Then, I get returns data from CRSP and analyst data from I/B/E/S. Finally, I obtain data on restatements from AuditAnalytics and AAER data from University of California, Berkeley's CFRM (Dechow et al. 2011) as extended by Bao et al. (2020). Since I need fundamentals data in each chapter, I always build my sample departing from Compustat data[16]. First, I apply some restrictions that are common through chapters. From the initial sample of 251,688 observations I only keep firms incorporated in the United States which lead to a sample of 191,848 firms for fiscal years between 1999 and 2018[17]. Then, I delete firms operating in the financial industry (*gics40*[18]) leading to 136,518 observations. Finally, I delete firms with missing data on sales or total assets leading to a base sample common to every chapter of 98,746 firm-year observations. Furthermore, for each chapter I apply additional restriction specific to the context of the study. Restrictions are detailed in each chapter.

In this chapter, I exclude very small firms (sales below 1 million USD or assets below 10 million USD), firms with missing data to calculate abnormal accruals, missing control variables, and observations in industries with less than 20 members in the same year. As shown in Table 2.1, this yields a full sample of 32,774 firm-year observations. The number of usable observations for tests involving AAER and market data is smaller due to additional required variables[19].

[16] Databases are consistently updated, even retrospectively. I download data at the same date for each database, to make replication easier. For this thesis, data were downloaded on 2020/03/30, except for the I/B/E/S data that comes from a prior date.

[17] I make this choice because I use the GICS classification which have been firstly issued in 1999.

[18] I also delete firms operating in industries *gics60* "Real Estate", after the introduction of this new industry in 2016.

[19] Bao et al.'s (2020) database (https://github.com/JarFraud/FraudDetection) includes AAERs disclosed until 2018-12-31, but I end the sample period in 2014 because it can take years before SEC investigation results are announced (Karpoff et al. 2017; Bao et al. 2020). For restatement tests, I end the sample period in 2017 because the delay between the publication of erreneous financial statements and their subsequent restatement is significantly shorter than AAERs (Karpoff et al. 2017) and our restatement data includes restatements announced until 2019-05-30.

TABLE 2.1: Sample selection

	Firm-year observations
Initial sample from Compustat with fiscal years between 1999 and 2018	251,688
Less firms incorporated outside of the US	(59,840)
Less firms in the financial industries	(55,330)
Less firms with missing sales or assets	(37,772)
Initial sample common to all chapters	98,746
Less sales under 1 million or assets under 10 million	(23,578)
Less missing data for accrual models	(24,241)
Less missing data for control variables	(14,652)
Less industries with less than 20 observations per year	(3,501)
Full sample	32,774

2.4. Results

2.4.1. Methodology results

Table 2.2 presents the distribution of industry classification misfits across GICS6 industries. Misfit firms (industry core firms) represent 12% (88%) of the sample. An average of 3.03 GICS6-SIC2 pairs have a large enough number of observations to be considered the core of any given GICS6 industry, and misfit firms are distributed across an average of 4.01 additional GICS6-SIC2 pairs. There are misfits in all GICS6 industries except Airlines and Electric Utilities.

Misfits firms are not uniformly distributed across GICS6 industries. For example, gics301010 "Food & Staples Retailing" includes only 2% of misfit firms, meaning that 98% of the firms inside this GICS6 industry share a GICS6-SIC2 combination that accounts for at least 5% of all firms . This result means that *gics301010* is composed of an important industry core group – including of 4.44 GICS6-SIC2 combinations on average – and misfit firms are not part of other significant clusters. On the other hand, *gics151010* "Chemicals" has an important proportion of misfit firms (20%).Its primary SIC2 equivalent is sic28 "Chemicals and Allied Products", which accounts for approximately 78% of this GICS6 industry, while the remaining 22% (misfits) are scattered across 10.12 unique SIC2 industries on average. Therefore, *gics151010* has a well-defined core with a relatively large proportion of misfit firms that are dispersed in several SIC2 industries.

TABLE 2.2: Industry statistics

GICS6 industry				Avg. no. unique SIC2 industries	
Code	**Description**	**N**	***% MISFIT***	**Core**	**Misfits**
101010	Energy Equipment & Services	961	20 %	3.59	8.35
101020	Oil, Gas & Consumable Fuels	2,594	11 %	4.88	6.82
151010	Chemicals	1,268	20 %	1.29	10.12
151030	Containers & Packaging	199	5 %	3.13	1.13
151040	Metals & Mining	781	13 %	3.88	4.12
151050	Paper & Forest Products	42	10 %	3.00	2.00
201010	Aerospace & Defence	870	18 %	4.71	6.76
201020	Building Products	417	9 %	5.41	2.29
201030	Construction & Engineering	395	8 %	3.50	2.06
201040	Electrical Equipment	915	10 %	3.06	4.06
201060	Machinery	1,837	13 %	3.18	6.47
201070	Trading Companies & Distributors	382	10 %	3.14	2.71
202010	Commercial Services & Supplies	1,813	29 %	5.39	16.67
202020	Professional Services	402	13 %	2.80	4.60
203020	Airlines	40	0 %	1.00	0.00
203040	Road & Rail	497	3 %	3.00	0.71
251010	Auto Components	607	12 %	3.65	4.12
252010	Household Durables	1,018	18 %	6.41	7.29
252020	Leisure Products	303	14 %	4.10	4.10
252030	Textiles, Apparel & Luxury Goods	870	13 %	4.94	5.18
253010	Hotels, Restaurants & Leisure	1,784	12 %	3.35	7.24
253020	Diversified Consumer Services	368	13 %	3.15	3.77
254010	Media	1,472	9 %	4.00	5.82
255020	Internet & Direct Marketing Reta	343	9 %	2.79	2.29
255030	Multiline Retail	146	3 %	1.50	0.83
255040	Specialty Retail	1,927	16%	4.11	10.17
301010	Food & Staples Retailing	467	2 %	4.44	0.63
302010	Beverages	61	8 %	1.00	1.67
302020	Food Products	973	7 %	2.00	2.88
303020	Personal Products	367	10 %	1.80	2.53
351010	Health Care Equipment & Supplies	2,138	6 %	2.00	6.06
351020	Health Care Providers & Services	1,803	22 %	4.35	9.82
351030	Health Care Technology	64	14 %	1.33	3.00
352010	Biotechnology	1,696	5 %	1.71	2.41
352020	Pharmaceuticals	812	5 %	1.00	1.88
352030	Life Sciences Tools & Services	359	2 %	4.25	0.50
501010	Diversified Telecommunication Se	537	4 %	1.13	1.19
501020	Wireless Telecommunication Servi	126	7 %	2.00	1.80
551010	Electric Utilities	577	0 %	1.00	0.00
551020	Gas Utilities	272	1 %	2.36	0.27
551030	Multi-Utilities	271	1 %	1.00	0.27
All industries		32,774	12 %	3.03	4.01

This table presents the distribution of core and misfit firms in the sample. For each GICS6 industry, the table shows the number of observations (*N*), the percentage of observations classified as industry misfits according to the procedure described in the text (*% MISFIT*), and the average number of SIC2 industries classified as core and misfit.

2.4.2. Univariate statistics

Table 2.3 Panel A presents descriptive statistics for misfits (*MISFIT*=1) and core firms (*MISFIT*=0). On average, misfit firms have higher absolute abnormal accruals and are more likely to have misstated financial statements. However, misfit firms are also smaller, more profitable and have lower fixed asset intensity (and a higher percentage of "soft" assets). Industry classification misfits also have higher sales volatility on average, although this does not translate into a higher volatility of operating cash flows. Table 2.3 Panel B also shows correlation coefficients for a subset of variables used in this chapter.

TABLE 2.3: Univariate analysis

Panel A: Descriptive statistics

	MISFIT=1			***MISFIT=0***			**Diff. (1-0)**			
Variable	***N***	**Mean**	**Median**	***N***	**Mean**	**Median**	**Mean**		**Median**	
$\|DACC\|_{NL}$	4,002	0.048	0.033	28,772	0.045	0.029	0.002	***	0.004	***
Restate	4,002	0.135	0.000	28,772	0.112	0.000	0.023	***	0.000	***
AAER	3,516	0.011	0.000	24,671	0.007	0.000	0.004	**	0.000	**
Size	4,002	6.038	5.995	28,772	6.453	6.386	-0.415	***	-0.390	***
ROA	4,002	0.015	0.040	28,772	0.000	0.036	0.015	***	0.004	***
TACC	4,002	-0.060	-0.050	28,772	-0.069	-0.056	0.008	***	0.005	***
CFO	4,002	0.075	0.084	28,772	0.070	0.086	0.005	**	-0.002	**
PPE	4,002	0.516	0.443	28,772	0.604	0.509	-0.089	***	-0.066	***
Peer_Shock	3,430	0.012	0.011	24,885	0.012	0.011	0.000		0.000	
Op_Shock	4,002	0.079	0.000	28,772	0.071	0.000	0.008	*	0.000	*
Idio_Shock	3,430	0.013	0.010	24,885	0.011	0.009	0.001	***	0.001	***
σ*(Sales)*	4,002	0.165	0.107	28,772	0.142	0.093	0.024	***	0.014	***
σ*(CFO)*	4,002	0.056	0.041	28,772	0.057	0.038	-0.001		0.003	**
BtoM	4,002	0.594	0.472	28,772	0.542	0.458	0.053	***	0.014	***
ChAuditor	4,002	0.074	0.000	28,772	0.071	0.000	0.004		0.000	
Issue3y	4,002	0.771	1.000	28,772	0.770	1.000	0.001		0.000	
BigN	4,002	0.773	1.000	28,772	0.791	1.000	-0.019	***	0.000	***
ΔReceivables	4,002	0.010	0.005	28,772	0.008	0.004	0.002	***	0.001	***
Δinventory	4,002	0.008	0.001	28,772	0.007	0.000	0.002	**	0.000	**
%_soft_assets	4,002	0.596	0.638	28,772	0.525	0.551	0.070	***	0.087	***
ΔCash_sales	4,002	0.141	0.059	28,772	0.139	0.058	0.002		0.001	
ΔROA	4,002	-0.005	0.000	28,772	-0.003	0.000	-0.002		-0.001	

TABLE 2.3 (CONTINUED): Univariate analysis

Panel B: Pearson correlations

	Var.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
1	*MISFIT*		0.02	0.02	0.01	-0.07	0.03	0.03	0.01	-0.07	0.00	0.01	0.04	0.05	-0.01	0.02	0.01	0.00	-0.01
2	$\lvert DACC\rvert_{NL}$			-0.01	-0.01	-0.23	-0.38	-0.33	-0.2	-0.06	0.14	0.07	0.22	0.10	0.26	-0.07	0.03	-0.07	-0.11
3	*Restate*				0.13	0.01	0.02	0.00	0.02	0.00	0.07	0.00	0.06	0.01	-0.02	0.01	0.01	0.03	0.02
4	*AAER*					0.03	0.01	0.02	0.00	-0.03	0.03	-0.01	0.02	0.02	-0.02	0.00	0.00	0.01	0.01
5	*Size*						0.25	0.07	0.24	0.11	-0.28	0.01	-0.48	-0.20	-0.41	-0.07	-0.11	0.32	0.44
6	*ROA*							0.50	0.77	0.08	-0.18	-0.02	-0.31	-0.01	-0.29	0.05	-0.03	0.00	0.05
7	*TACC*								-0.15	-0.21	-0.06	0.07	-0.12	0.02	-0.07	0.08	-0.03	-0.02	0.01
8	*CFO*									0.25	-0.16	-0.08	-0.27	-0.02	-0.28	-0.01	-0.01	0.01	0.06
9	*PPE*										-0.23	-0.05	-0.11	-0.13	-0.17	0.01	0.00	0.15	-0.01
10	*Peer_Shock*											-0.01	0.46	0.02	0.22	-0.03	0.07	-0.12	0.06
11	*Op_Shock*												0.01	-0.01	0.02	0.00	0.00	0.02	-0.02
12	*Idio_Shock*													0.19	0.33	-0.01	0.09	-0.11	-0.14
13	σ*(Sales)*														0.31	0.01	0.03	-0.08	-0.12
14	σ*(CFO)*															-0.02	0.03	-0.19	-0.15
15	*BtoM*																0.01	-0.03	-0.08
16	*ChAuditor*																	-0.01	-0.15
17	*Issue3y*																		0.13
18	*BigN*																		

All variables are defined in the Appendix. This table presents descriptive statistics and a correlation analysis. Panel A presents the mean and median values for industry misfits (MISFIT=1) and core (MISFIT=0) observations, along with the differences in means and medians between both groups (Diff.). A *** (**; *) indicates a significant difference in means (medians) at the 1% (5%; 10%) level using a t-test (a Wilcoxon signed rank test). Panel B presents Pearson correlation coefficients between a subset of variables for the full sample.

2.4.3. Accruals

Table 2.4 reports results of the regression tests of the association between absolute abnormal accruals and *MISFIT* (H1). Model (1) is based on Eq. (2.2). The coefficient on *MISFIT* is positive (0.0027) and significant at the 1% level. Consistent with H1, misfit firms exhibit higher absolute abnormal accruals on average. This result is consistent with the argument that misfit firms are heterogenous to their industry, which generates measurement error in the first-stage accrual prediction model.

In Models (2) and (3), I examine whether the positive association between *MISFIT* and accruals holds when shock variables are included (Owens et al. 2017), and whether misfit and core firms' absolute abnormal accruals are differently affected by shocks to the accrual generating process experienced by other peer firms.. In Model (2), I add three shock variables in Owens et al. (2017) to Model (1). The coefficient on *MISFIT* remains positive and significant at the 1% level, consistent with H1. The coefficients on $Peer\_Shock_{t}$, Op_Shock_{t} and $Idio\_Shock_{t-1}$ are all positive and significant at the 1% level, consistent with Owens et al. (2017) and indicating that both own-firm and peer-firm shocks affect the estimation of accruals. Model (3) is based on Eq. (2.3). The model includes interaction terms between $Peer\_Shock_{t-1}$ and $MISFIT_{t}$ to allow the peer-shock effect to differ for misfits and core firms. The coefficient on $Peer\_Shock_{t-1}$ is positive and significant at the 1% level, indicating that all firms' accruals are affected by peer shocks. However, the coefficient on the interaction term is negative and significant at the 1% level, indicating that industry classification misfits are less affected than core firms by the peer shock effect documented by Owens et al. (2017). Finally, the coefficient on $MISFIT_{t}$ remains positive and significant at the 1% level and is numerically larger in Model (3), suggesting that after controlling for the effect of peer shocks on accrual estimation, industry classification misfits still have significantly larger absolute abnormal accruals than industry core firms.

TABLE 2.4: Industry misfits and absolute abnormal accruals

Model	**(1)**			**(2)**			**(3)**		
Variable	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
MISFIT	**0.0027**	***	**2.7**	**0.0034**	***	**3.3**	**0.0080**	***	**3.2**
Peer_Shock$_{t-1}$				**0.9479**	***	**11.3**	**0.9886**	***	**11.5**
Peer_Shock$_{t-1}$****MISFIT***							**-0.3796**	**	**-2.1**
Idio_Shock$_{t-1}$				0.2253	***	5.2	0.2256	***	5.2
Op_Shock				0.0135	***	10.2	0.0135	***	10.2
(ΔSALES-ΔREC)	0.0034	**	2.2	0.0024		1.4	0.0024		1.4
PPE	-0.0059	***	-6.0	-0.0034	***	-3.2	-0.0034	***	-3.1
CFO$_{t-1}$	0.0194	***	3.9	0.0213	***	3.9	0.0214	***	3.9
CFO$_{t}$	0.0784	***	12.6	0.0756	***	11.1	0.0756	***	11.1
CFO$_{t+1}$	0.0162	***	4.1	0.0212	***	4.7	0.0212	***	4.7
DCF$_{t}$	0.0153	***	10.6	0.0132	***	8.5	0.0132	***	8.5
DCF$_{t}$**CFO*$_{t}$	0.0631	***	7.0	0.0792	***	8.0	0.0794	***	8.0
ROA	-0.1754	***	-24.3	-0.1840	***	-23.3	-0.1840	***	-23.3
Size	-0.0021	***	-9.5	-0.0014	***	-5.9	-0.0014	***	-5.8
σ(CFO)	0.1228	***	14.7	0.1121	***	12.5	0.1117	***	12.4
σ(Sales)	0.0077	***	3.5	0.0089	***	3.8	0.0089	***	3.9
BtoM	-0.0026	***	-7.4	-0.0026	***	-7.5	-0.0026	***	-7.5
ChAuditor	0.0010		0.9	0.0003		0.3	0.0003		0.3
Issue3y	-0.0017	**	-2.1	-0.0016	*	-1.9	-0.0016	*	-1.9
BigN	-0.0033	***	-3.2	-0.0029	***	-2.7	-0.0029	***	-2.7
Intercept	0.0544	***	23.5	0.0347	***	13.3	0.0341	***	13.0
Year fixed effects	Incl.			Incl.			Incl.		
GICS6 Industry fixed effects	Not incl.			Not incl.			Not incl.		
Adjusted R^2	0.24			0.26			0.26		
N	32,774			28,315			28,315		

All variables are defined in the Appendix. This table reports coefficients estimates (Coeff.), statistical significance (Sig.) and t-statistics (t.) from a regression of absolute discretionary accruals (|*DACC*|$_{NL}$) on *MISFIT* and control variables. In the Sig. column, a *** (**; *) indicates that the coefficient is different from zero at the 1% (5%; 10%) level. All standard errors are clustered by firm.

2.4.4. Misstatements

Table 2.5 shows results of regressions of misstatement proxies (*Restate*$_t$ and *AAER*$_t$) on absolute abnormal accruals and other determinants. For both proxies, I estimate three models. First, I estimate a benchmark model where the variable of interest is $|DACC|_{NL}$. Second, I include an interaction between this variable and a dummy for misfit firms. A significant negative coefficient on $|DACC|_{NL}*MISFIT_t$ would be consistent with the hypothesis that abnormal accruals are less associated with misstatements for misfit firms than for industry core firms (H2). Third, I estimate the benchmark model using a sample restricted to core firms only. I estimate models with logistic regression (Table 2.5 Panel A and C) and with OLS (Table 2.5 Panel B and D) to avoid issues with the interpretation of interaction terms in logistic regressions (Ai and Norton 2003); I only discuss the former because both sets of results are qualitatively similar on all aspects.

TABLE 2.5: Industry misfits, accruals and the prediction of misstatements

Panel A: Restatements - Logistic regressions

Sample		***Full***			***Full***			***No misfits***	
	Coeff.	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
$\|DACC\|_{NL}$	**0.9009**	**	**2.1**	**1.3094**	***	**2.8**	**1.3077**	***	**2.8**
$\|DACC\|_{NL}$*MISFIT				**-3.1344**	**	**-2.4**			
MISFIT				0.2997	***	3.0			
ΔReceivables	-0.6725		-1.5	-0.6836	*	-1.6	-0.5849		-1.2
Δinventory	0.2031		0.4	0.2028		0.4	0.0463		0.1
%_soft_assets	0.4370	***	2.9	0.4228	***	2.8	0.4074	**	2.4
ΔCash_sales	0.0810		1.2	0.0791		1.2	0.1311		1.8
ΔROA	0.0172		0.1	0.0120		0.1	-0.0353		-0.3
Issue3y	0.2367	***	2.8	0.2307	***	2.7	0.2084	**	2.4
Size	0.0182		1.0	0.0211		1.1	0.0140		0.7
σ(Sales)	0.1356		0.8	0.1359		0.8	0.0958		0.5
BtoM	0.0234		1.1	0.0231		1.1	0.0266		1.3
BigN	-0.0089		-0.1	-0.0160		-0.2	-0.0015		0.0
Intercept	-2.8472	***	-18.1	-2.8804	***	-18.2	-2.8340	***	-16.6
Year fixed effects	Incl.			Incl.			Incl.		
GICS6 industry fixed effects	Incl.			Incl.			Incl.		
Max rescaled R^2	0.06			0.06			0.07		
Area under ROC	0.65			0.65			0.66		
N	32,774			32,774			28,772		

Panel B: Restatements - OLS regressions

Sample	***Full***			***Full***			***No misfit***		
	Coeff.	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
$|DACC|_{NL}$	**0.0904**	**	**2.2**	**0.1289**	***	**2.9**	**0.1270**	***	**2.9**
$|DACC|_{NL}$*MISFIT				**-0.3131**	***	**-2.6**			
MISFIT				0.0318	***	2.8			
ΔReceivables	-0.0713	*	-1.6	-0.0723	*	-1.6	-0.0623		-1.4
Δinventory	0.0267		0.5	0.0253		0.5	0.0147		0.3
%_soft_assets	0.0407	***	2.7	0.0395	***	2.7	0.0364	**	2.3
ΔCash_sales	0.0077		1.3	0.0076		1.3	0.0116	**	1.9
ΔROA	0.0018		0.2	0.0010		0.1	-0.0032		-0.3
Issue3y	0.0216	***	2.9	0.0210	***	2.9	0.0190	**	2.5
Size	0.0018		1.0	0.0021		1.1	0.0014		0.7
σ*(Sales)*	0.0158		0.9	0.0156		0.9	0.0125		0.7
BtoM	0.0026		1.2	0.0026		1.2	0.0029		1.3
BigN	-0.0014		-0.2	-0.0022		-0.3	-0.0006		-0.1
Intercept	-0.0486		-1.3	-0.0503		-1.4	-0.0354		-1.0
Year fixed effects	Incl.			Incl.			Incl.		
GICS6 industry fixed effects	Incl.			Incl.			Incl.		
Adjusted R^2	0.03			0.03			0.04		
N	32,774			32,774			28,772		

Panel C: AAER – Logistic regressions

Sample	***Full***			***Full***			***No misfits***		
	Coeff.	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
$\boldsymbol{|DACC|_{NL}}$	**-0.8757**		**-0.6**	**-0.9064**		**-0.5**	**-0.7627**		**-0.5**
$\boldsymbol{|DACC|_{NL}*MISFIT}$				**-0.0004**		**0.0**			
MISFIT				0.2286		0.7			
ΔReceivables	1.4601		1.0	1.4858		1.0	1.1113		0.7
Δinventory	-0.1846		-0.1	-0.1522		-0.1	-0.8338		-0.3
%_soft_assets	1.6898	**	2.4	1.6590	**	2.3	2.1759	***	2.7
ΔCash_sales	0.6494	**	2.4	0.6433	**	2.4	0.7094	**	2.3
ΔROA	-0.5964		-1.0	-0.5927		-1.0	-0.5010		-0.9
Issue3y	0.4791		1.2	0.4687		1.2	0.5056		1.2
Size	0.3209	***	4.7	0.3231	***	4.7	0.3417	***	4.3
σ(Sales)	1.4545	***	2.6	1.4362	***	2.6	1.3929	**	2.1
BtoM	0.0283		0.6	0.0267		0.6	0.0120		0.3
BigN	-0.7319		-2.2	-0.7347	**	-2.2	-0.6311	*	-1.6
Intercept	-11.4356	***	-14.8	-11.4392	***	-14.8	-12.2645	***	-13.8
Year fixed effects	Incl.			Incl.			Incl.		
GICS6 Industry fixed effects	Incl.			Incl.			Incl.		
Max rescaled R^2	0.13			0.13			0.15		
Area under ROC	0.81			0.82			0.82		
N	28,187			28,187			24,671		

Panel D: AAER – OLS Regressions

Sample	***Full***			***Full***			***No misfit***		
	Coeff.	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
$\lvert DACC\rvert_{NL}$	**-0.0024**		**-0.3**	**-0.0016**		**-0.2**	**-0.0004**		**0.0**
$\lvert DACC\rvert_{NL}$\*MISFIT				**-0.0084**		**-0.2**			
MISFIT				0.0030		0.8			
ΔReceivables	0.0170		1.2	0.0170		1.2	0.0125		0.8
Δinventory	0.0031		0.1	0.0030		0.1	0.0009		0.0
%_soft_assets	0.0102	**	2.3	0.0099	**	2.2	0.0118	***	2.7
ΔCash_sales	0.0049	**	2.0	0.0049	**	2.0	0.0043	*	1.8
ΔROA	-0.0035		-1.4	-0.0035		-1.4	-0.0027		-1.2
Issue3y	0.0023		1.3	0.0022		1.3	0.0023		1.4
Size	0.0027	***	3.8	0.0027	***	3.9	0.0027	***	3.5
σ(Sales)	0.0155	**	2.3	0.0154	**	2.3	0.0135	*	1.9
BtoM	0.0002		0.8	0.0002		0.8	0.0001		0.4
BigN	-0.0058	**	-2.4	-0.0059	**	-2.4	-0.0049	*	-1.9
Intercept	-0.0299	***	-4.1	-0.0301	***	-4.1	-0.0299	***	-3.8
Year fixed effects	Incl.			Incl.			Incl.		
GICS6 Industry fixed effects	Incl.			Incl.			Incl.		
Adjusted R^2	0.01			0.01			0.01		
N	28,187			28,187			24,671		

All variables are defined in the Appendix. Panel A and C (Panel B and D) reports coefficients estimates (Coeff.) and statistical significance (Sig.) from logistic (OLS) regressions of two proxies for future misstatements (Restate and AAER) on absolute discretionary accruals ($|\text{DACC}|_{\text{NL}}$), on MISFIT, an interaction between $|\text{DACC}|_{\text{NL}}$ and MISFIT, and control variables. In the Sig. column, a *** (**; *) indicates that the coefficient is different from zero at the 1% (5%; 10%) level (two-tailed). All standard errors are clustered by firm.

In Table 2.5 Panel A, when $Restate_t$ is the dependent variable, the coefficient on $|DACC|_{NL}$ is positive (0.9009) and significant at the 5% level, suggesting that firms with higher absolute abnormal accruals subsequently restate their annual financial statements with a greater frequency than firms with lower accruals. I observe two notable changes when the model includes an interaction term to enable the coefficient on $|DACC|_{NL}$ to vary between misfits and core firms. First, the coefficient on $|DACC|_{NL}*MISFIT_t$ is negative and significant at the 1% level, consistent with the hypothesis that the association between abnormal accruals and misstatements is smaller for misfit firms (H2.2). Second, in contrast, the coefficient on $|DACC|_{NL}$ is numerically larger (1.3094) and is significant at the 1% level, a result that is also obtained in Model (3), when the sample is restricted to industry core firms (1.3077). Taken together, these results suggest that for industry core firms, absolute abnormal accruals are incrementally useful to predict future restatements. However, for misfit firms, the measurement error in abnormal accruals is so large that there is no statistical association between absolute abnormal accruals and restatements, two widely used proxies for financial reporting quality.

As for tests with AAER as the dependent variable (Table 5 Panel C), results indicate no association between $|DACC|_{NL}$ and $AAER_t$ in any of three models, suggesting that absolute abnormal accruals are not incremental predictors of future SEC enforcement actions, whether for industry core or misfit firms. AAERs are relatively infrequent (less than 1% of observations), which could lead to a lack of power of statistical tests. Also, the model specification could suffer from the inclusion of control variables that are highly correlated with both accruals and *AAER* such as *Soft_assets* or *ΔCash_sales*. As an additional test, I alternatively and simultaneously exclude these two variables from the regression. Results remain unaffected by these changes as the coefficient on $|DACC|_{NL}$ is still insignificant.

2.4.5. Return comovement

Table 2.6 presents results of tests of H2.3 on the association between firm-specific stock returns and industry returns for misfits and core firms. Model (1) is based on Eq. (2.5); Model (2) (Model (3)) restricts the sample to misfit (core) firms only and therefore exclude all regressors that include *MISFIT*. In Model (1), as expected, the coefficient on $INDRET_t$ is positive and (highly) significant (t-stat = 131.7) confirming that firm returns are strongly driven by industry news. However, the coefficient on $IND\_RET_t*MISFIT_t$ is negative (-0.0914) and significant at the 1% level (t-stat = -5.7). This is consistent with H2.3 and suggests that misfit firms have a smaller industry beta than industry core firms. In other words, misfit firms' stock returns are less affected by industry news than peers in the industry core, consistent with their misfit status. Other interaction terms are not significant, indicating that this weaker contemporaneous association is not offset by a greater delayed incorporation of industry news in stock prices or by a different autoregressive structure of stock returns for misfit firms. Results in Models (2) and (3) tell a qualitatively similar story. Control variable coefficients and significance are in accordance with previous literature (e.g. Cohen and Lou 2012). Overall, my results support H2.3.

TABLE 2.6: Industry misfits and return comovement

Model	**(1)**			**(2)**			**(3)**		
Sample	**Full**			**Misfits**			**No misfits**		
Variable	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
$INDRET_t$	**0.8772**	***	**131.7**	**0.7857**	***	**52.3**	**0.8772**	***	**131.8**
$INDRET_t$*MISFIT	**-0.0914**	***	**-5.7**						
$INDRET_{t-1}$	0.0693	***	33.3	0.0691	***	13.2	0.0693	***	33.4
*$INDRET_{t-1}$*MISFIT*	-0.0002		-0.0						
RET_{t-1}	-0.0412	***	-29.1	-0.0405	***	-11.8	-0.0412	***	-29.1
*RET_{t-1}*MISFIT*	0.0009		0.3						
MISFIT	0.0066	*	1.7						
Size	-0.0024	***	-2.8	-0.0009		-0.3	-0.0029	***	-3.1
BtoM	-0.0100	**	-2.1	-0.0229		-1.6	-0.0085	*	-1.8
Turnover	0.0151	***	9.5	0.0096	**	2.1	0.0158	***	9.6
Intercept	-0.0000		-0.0	-0.0015	***	-3.5	0.0000		0.1
Year fixed effects	Incl.			Incl.			Incl.		
GICS6 Industry fixed effects	Incl.			Incl.			Incl.		
Adjusted R^2	0.18			0.15			0.19		
N	7,381,455			892,043			6,489,412		

All variables are defined in the Appendix. This table reports coefficients estimates (Coeff.), statistical significance (Sig.) and t-statistics (t.) from a regression of daily returns (*RET_t*) on *MISFIT*, current and lagged industry returns (*$INDRET_t$* and *$INDRET_{t-1}$*) and control variables. In the Sig. column, a *** (**; *) indicates that the coefficient is different from zero at the 1% (5%; 10%) level (two-tailed). Coefficients on non-returns variables (*Size, BtoM, Turnover, MISFIT*) are multiplied by 100. All standard errors are clustered by firm.

2.5. Additional analysis

2.5.1. Other accruals models

To make sure that my results are not an artifact of a specific model, I use three other accrual prediction models to estimate abnormal accruals. All models are estimated by GICS6 industry and by year:

$$TACC_t = \alpha + \beta_1(1/AT_{t-1}) + \beta_2(\Delta SALE_t - \Delta RECT_t) + \beta_3 PPEGT_{t-1} + \varepsilon_t \qquad (2.6;\ MODJ)$$

$$TACC_t = \alpha + \beta_1 CFO_{t-1} + \beta_2 CFO_t + \beta_3 CFO_{t+1} + \varepsilon_t \qquad (2.7;\ DD)$$

$$TACC_t = \alpha + \beta_1 CFO_{t-1} + \beta_2 CFO_t + \beta_3 CFO_{t+1} + \beta_4(\Delta SALE_t - \Delta RECT_t) \qquad (2.8;\ MN)$$

$$+ \beta_5 PPEGT_{t-1} + \varepsilon_t$$

The dependent variable in all models is total accruals ($TACC_t$). The models are based on Dechow, Sloan and Sweeney (1995; modified Jones i.e., *MODJ*), Dechow and Dichev (2002; *DD*) and McNichols (2002; *MN*). As with the nonlinear model used in main tests, the absolute value of the residual yields a proxy of absolute abnormal accruals: $|DACC|_{MODJ,}$ $|DACC|_{DD,}$ $|DACC|_{MN}$. For each proxy, I replicate my main analyses and report results in Table 2.7. Table 2.7 Panel A shows that misfits have higher absolute abnormal accruals than industry core firms, regardless of the accrual prediction model used in the first stage. Table 2.7 Panel B confirms this in a regression setting, as the coefficient on $MISFIT_t$ is positive (0.0035, 0.0032 and 0.0029 respectively) and significant at the 1% level with all three alternative models. Thus, misfit firms experience higher absolute abnormal accruals, a finding that I attribute to a poorer fit of accrual prediction models for these firms.

Table 2.7 Panel C shows logistic regression results with *Restate* and *AAER* as a dependent variable. I only report the coefficient on $|DACC|_{x,}$ where x indexes the model selected. The results are consistent with earlier findings, as absolute abnormal accruals are positively associated with future restatements for industry core firms but not for misfit firms, while absolute abnormal accruals have no incremental explanatory power for the prediction of AAERs for either group.

TABLE 2.7: Other accrual models

Panel A: Descriptive statistics

	MISFIT=1			***MISFIT=0***			**Diff. (1-0)**			
Variable	***N***	**Mean**	**Median**	***N***	**Mean**	**Median**	**Mean**		**Median**	
$\|DACC\|_{NL}$	4002	0.048	0.033	28772	0.045	0.029	**0.002**	***	**0.004**	***
$\|DACC\|_{MODJ}$	4002	0.062	0.041	28772	0.057	0.037	**0.005**	***	**0.004**	***
$\|DACC\|_{DD1}$	4002	0.055	0.037	28772	0.051	0.032	**0.004**	***	**0.005**	***
$\|DACC\|_{DD3}$	4002	0.051	0.035	28772	0.048	0.031	**0.003**	***	**0.004**	***

This table presents descriptive statistics and a correlation analysis. Variable definitions are in the Appendix. Panel A presents the mean and median values for industry misfits (*MISFIT*=1) and core (*MISFIT*=0) observations, along with the differences in means and medians between both groups (Diff.). A *** (**; *) indicates a significant difference at the 1% (5%; 10%) level using a t-test (a Wilcoxon signed rank test). Panel B presents Pearson correlation coefficients between a subset of variables

Panel B: Industry misfits and absolute abnormal accruals

Model	**(1)**			**(2)**			**(3)**		
Dep. variable	$\|DACC\|_{MODJ}$			$\|DACC\|_{DD1}$			$\|DACC\|_{DD3}$		
Variable	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
MISFIT	**0.0035**	***	**2.95**	**0.0032**	***	**2.8**	**0.0029**	***	**2.9**
Control variables	Incl.			Incl.			Incl.		
Adjusted R^2	0.27			0.28			0.28		
N	32,774			32,774			32,774		

All variables are defined in the Appendix. This table reports coefficients estimates (Coeff.), statistical significance (Sig.) and t-statistics (t.) from a regression of absolute abnormal accruals using alternative accrual models on *MISFIT* and control variables. In the Sig. column, a *** (**; *) indicates that the coefficient is different from zero at the 1% (5%; 10%) level (two-tailed). All standard errors are clustered by firm.

Panel C: Accruals and the prediction of misstatements (logistic regressions)

Dep. variable	***Restate***		***Restate***		***AAER***		***AAER***	
Sample	**Misfits**		**No misfits**		**Misfits**		**No misfits**	
	Coeff.	**t.**	**Coeff.**	**t.**	**Coeff.**	**t.**	**Coeff.**	**t.**
$\|DACC\|_{MODJ}$	**-0.5530**	**-0.6**	**0.9357**	**2.4**	**-3.4150**	**-0.9**	**-0.1375**	**-0.1**
Control variables	Incl.		Incl.		Incl.		Incl.	
$\|DACC\|_{DD1}$	**-0.4566**	**-0.4**	**0.8931**	**2.1**	**-6.1836**	**-1.6**	**-1.8126**	**-1.0**
Control variables	Incl.		Incl.		Incl.		Incl.	
$\|DACC\|_{DD3}$	**-0.7157**	**-0.6**	**1.4066**	**3.1**	**-2.8586**	**-0.6**	**-0.3835**	**-0.2**
Control variables	Incl.		Incl.		Incl.		Incl.	

All variables are defined in the Appendix. This table reports coefficients estimates (Coeff.) and statistical significance (Sig.) from logistic regressions of two proxies for future misstatements (*Restate* and *AAER*) on absolute abnormal accruals using various models for the full sample and for industry core firms only ("No misfits" sample). All standard errors are clustered by firm.

2.5.2. Alternative definitions of misfit

In my methodology, I subjectively define misfits as firms whose GICS6-SIC2 combination constitutes less than five percent of all firms with the same GICS6 for that year. In order to show that my results are not driven by this arbitrary choice, I provide alternative definitions of misfit firms based on two different thresholds. First, I classify as misfits all firms that do not belong to the most frequent SIC2 within a given GICS6 industry. Second, I classify as misfits all firms in GICS6-SIC2 combinations that include less than five observations. To evaluate the effect of these alternative definitions, I estimate Eq. (2.2) and (2.3) using the revised *MISFIT* values. The results are in Table 2.8. In models (2) and (4), the coefficient on MISFIT is positive and significant at the 1% level, indicating that earlier results on H1 are not restricted to the specific definition of misfit firms used in main tests. In addition, the interaction between *Peer_Shock* and *MISFIT* yields to a negative and significant coefficient with both alternative definitions, indicating that core firms are more affected by other firm shocks than misfits. However, results in models (1) and (3) are surprisingly insignificant.

TABLE 2.8: Alternative misfit definitions and absolute abnormal accruals

Definition of *MISFIT*				**Not the most frequent SIC2**						**Less than 5 observations in GICS6-SIC2 pair**		
Model	**(1)**			**(2)**			**(3)**			**(4)**		
Variable	**Coeff.**	**Sign.**	**t.**	**Coeff.**	**Sign.**	**t.**	**Coeff.**	**Sign.**	**t.**	**Coeff.**	**Sign.**	**t.**
MISFIT	**0.0009**		**1.3**	**0.0089**	***	**6.0**	**-0.0002**		**-0.2**	**0.0053**	***	**2.8**
***Peer_Shock*$_{t-1}$**				**1.1595**	***	**12.3**				**1.0018**	***	**11.3**
Peer_Shock*$_{t-1}$**MISFIT				**-0.6335**	***	**-5.6**				**-0.3784**	**	**-2.5**
Idio_Shock$_{t-1}$				0.2241	***	5.1				0.2254	***	5.2
Op_Shock				0.0136	***	10.3				0.0136	***	10.3
(ΔSALES-ΔREC)	0.0034	**	2.2	0.0024		1.5	0.0034	**	2.2	0.0024		1.4
PPE	-0.0060	***	-6.0	-0.0032	***	-3.0	-0.0061	***	-6.1	-0.0034	***	-3.2
CFO$_{t-1}$	0.0195	***	3.9	0.0216	***	4.0	0.0196	***	3.9	0.0214	***	4.0
CFO$_{t}$	0.0785	***	12.6	0.0756	***	11.0	0.0782	***	12.6	0.0755	***	11.0
CFO$_{t+1}$	0.0163	***	4.1	0.0213	***	4.8	0.0163	***	4.1	0.0212	***	4.7
DCF$_{t}$	0.0154	***	10.6	0.0133	***	8.6	0.0153	***	10.6	0.0132	***	8.5
DCF$_{t}$**CFO*$_{t}$	0.0629	***	6.9	0.0798	***	8.0	0.0635	***	7.0	0.0795	***	8.0
ROA	-0.1753	***	-24.3	-0.1840	***	-23.3	-0.1752	***	-24.3	-0.1838	***	-23.2
Size	-0.0021	***	-9.7	-0.0014	***	-6.0	-0.0022	***	-9.8	-0.0014	***	-6.0
σ(CFO)	0.1227	***	14.7	0.1113	***	12.4	0.1221	***	14.6	0.1112	***	12.4
σ(Sales)	0.0076	***	3.5	0.0086	***	3.7	0.0079	***	3.6	0.0092	***	4.0
BtoM	-0.0026	***	-7.5	-0.0026	***	-7.4	-0.0026	***	-7.4	-0.0026	***	-7.5
ChAuditor	0.0009		0.9	0.0003		0.2	0.0010		0.9	0.0003		0.3
Issue3y	-0.0016	**	-2.1	-0.0016	*	-1.9	-0.0016	**	-2.0	-0.0016	*	-1.8
BigN	-0.0033	***	-3.2	-0.0029	***	-2.7	-0.0033	***	-3.2	-0.0029	***	-2.6
Intercept	0.0545	***	23.2	0.0318	***	11.7	0.0551	***	23.6	0.0344	***	13.0
Year fixed effects	Incl.			Incl.			Incl.			Incl.		
GICS6 Industry fixed effects	Not incl.			Not incl.			Not incl.			Not incl.		
Adjusted R^{2}	0.26			0.26			0.26			0.26		
N	28,315			28,315			28,315			28,315		

All variables are defined in the Appendix. This table reports coefficients estimates (Coeff.) and statistical significance (Sig.) from a regression of absolute abnormal accruals ($|DACC|_{NL}$) on MISFIT and control variables, where MISFIT is defined differently than in main tests. In models (1) and (2), only firms with the most frequent GICS6-SIC2 pair in a given GICS6 for that year are considered in the industry core and all others are considered misfits. In models (3) and (4), all firms in GICS6-SIC2 pairs containing at least five observations are considered in the industry core and all others are considered misfits. In the Sig. column, a *** (**; *) indicates that the coefficient is different from zero at the 1% (5%; 10%) level (two-tailed). All standard errors are clustered by firm.

2.5.3. Matched sample

A potential concern regarding my methodology is that it may create two groups of firms (misfit and industry core firms) that are dissimilar on certain dimensions that could confound the interpretation of earlier results. For example, Table 2.3 Panel A shows that misfit firms are smaller than industry core firms. Considering that firms of different size may not be comparable (Ecker et al. 2013) and that smaller firms generally experience a poorer fit in accruals models, I redo my analysis using a matched sample. I match each misfit firm with an industry core firm of comparable size in the same GICS6 industry and in the same year. This procedure yields to a matched sample of 7,578 firms.

Table 2.9 provides the empirical results using this sample. All results are qualitatively similar to the main results. Specifically, the mean and median $|DACC|_{NL}$ are significantly higher for misfit firms than matched industry core firms in univariate and regression tests (respectively, Table 9 Panel A and B), the positive association between absolute abnormal accruals and future restatements is lower for misfits than matched core firms (Table 9 Panel C), and the association between industry news and misfit firms' stock returns is weaker than for industry core firms' stock returns. I conclude that our main results are not driven by differences in average firm size between misfits and industry core firms.

TABLE 2.9: Matched sample

Panel A: Descriptive statistics

	MISFIT=1			***MISFIT=0***			**Diff. (1-0)**			
Variable	***N***	**Mean**	**Median**	***N***	**Mean**	**Median**	**Mean**		**Median**	
$\|DACC\|_{NL}$	3789	0.0480	0.0327	3789	0.0460	0.0301	**0.0021**	*	**0.0026**	***
$\|DACC\|_{MODJ}$	3789	0.0611	0.0407	3789	0.0584	0.0391	**0.0027**	*	**0.0016**	
$\|DACC\|_{DD}$	3789	0.0554	0.0377	3789	0.0522	0.0342	**0.0032**	**	**0.0035**	***
$\|DACC\|_{MN}$	3789	0.0510	0.0344	3789	0.0489	0.0327	**0.0020**	*	**0.0017**	***

Panel B: Industry misfits and absolute abnormal accruals

Model	**(1)**			**(2)**			**(3)**		
Dep. variable	$\mathbf{|DACC|_{NL}}$			$\mathbf{|DACC|_{MODJ}}$			$\mathbf{|DACC|_{MN}}$		
Variable	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
MISFIT	**0.0030**	**	**2.5**	**0.0036**	**	**2.5**	**0.0029**	***	**2.4**
Control variables	Incl.			Incl.					
Adjusted R^2	0.26			0.27			0.29		
N	7,578			7,578			7,578		

Panel C: Accruals and the prediction of misstatements (logistic regressions)

Dep. variable	**Restate**		**Restate**		**AAER**		**AAER**	
	Coeff.	**t.**	**Coeff.**	**t.**	**Coeff.**	**t.**	**Coeff.**	**t.**
$\mathbf{|DACC|_{NL}}$	**0.6252**	**0.8**	**2.4690**	**2.4**	**1.4778**	**0.5**	**3.1534**	**1.0**
$\mathbf{|DACC|_{NL}*MISFIT}$			**-3.7452**	**-2.4**			**-3.3748**	**-0.6**
MISFIT			0.3720	3.0			0.1749	0.4
Control variables	Incl.		Incl.		Incl.		Incl.	

Panel D: Industry misfits and return comovement

Model	**(1)**		
Sample	**Full**		
Variable	**Coeff.**	**Sig.**	**t.**
$INDRET_t$	0.8652	***	65.3
$\mathbf{INDRET_t*MISFIT}$	**-0.0747**	***	**-3.8**
Controls	Incl.		
Year fixed effects	Incl.		
GICS6 Industry fixed effects	Incl.		
Adjusted R^2	0.17		
N	1,675,408		

All variables are defined in the Appendix. This table reports descriptive statistics and coefficient results from a sample of misfits and industry core firms matched on GICS6 industry, year and size. Panel A presents descriptive statistics for absolute abnormal accruals. Panel B presents the replication of the result from Table 2.7 Panel B. Panel C replicates the results from Table 2.5 Panel A and C (models (1) and (2)), while Panel D presents the replication of Table 2.6 model (1).

2.6. Summary

In contrast with prior studies (e.g. Bhojraj et al. 2003, Hrazdil and Scott 2013) that consider the different industry classification schemes as substitutes, this chapter provides evidence that there is value in regarding multiple schemes as complements. I develop a simple methodology that relies on the convergence between the two main classifications (i.e., GICS6 and SIC2) and use it to identify firms that are sources of intra-industry heterogeneity. I refer to these firms as industry classification misfit firms and study the consequences of their inclusion into peer groups.

First, I provide evidence regarding accruals models. Usually, accruals models are estimated by industry and year in order to identify firms that deviate from their benchmark, an implicit assumption being that industries form a homogeneous group of firms that share the same accrual-generating pattern. Since misfit firms represent sources of intra-industry heterogeneity, I posit that they will suffer from a poorer fit for these models. My results support this hypothesis, as misfit firms experience higher absolute abnormal accruals. Then, I test the consequences of this measurement error for the predictions of misstatements. I show that absolute abnormal accruals are less associated with future restatements for misfit firms than for industry core firms, consistent with the argument that absolute abnormal accruals are a noisier proxy of financial reporting quality for misfit firms. Finally, I focus on the association between industry stock returns and firm-specific stock returns. I show that misfit firms' stock returns are less affected by industry news than industry core firms' stock returns, consistent with the argument that they are economically heterogeneous to their industry peers.

Results from robustness tests are generally consistent with main results. First, in main analysis, I estimate abnormal accruals with the Ball and Shivakumar (2006) nonlinear model. To make sure that my results are not driven by this choice, I use three other accrual models and obtain qualitatively similar results. Also, my methodology requires arbitrary choices regarding the definition of misfit firms. In supplementary analyses, I show alternative specifications do not affect substantially our results. Finally, results from a subsample of misfits and size-matched industry core firms remain qualitatively similar.

The main limitation of my new methodology is a direct consequence of its operational simplicity. While the extent to which a firm deviates from the industry mean may be best measured on a continuum, my method to identify heterogeneous firms is based on a simple combination of industry classification codes and yields a binary outcome (e.g., misfit vs. core). I intentionally designed this method to show complementarity between industry classifications in a simple setting, but my methodology does not provide an answer regarding an optimal peer selection method.

3. Differentiated firms

3.1. Introduction

Many reasons could lead firms to voluntarily differentiate themselves from their peers in competitive markets. For example, for strategic purposes, firms may use product differentiation to gain a competitive advantage. Even though the evidence is clear about the benefits of differentiation, firms still imitate each other (Lieberman and Asaba 2006), suggesting there may be some cost associated with the pursuit of a differentiation strategy. In this chapter, I build on previous literature (Foucault and Frésard 2018) and argue that the lack of benchmarking for differentiated firms can create informational costs which could potentially reduce the benefits of differentiation.

According to both the IFRS and US GAAP conceptual frameworks, a fundamental qualitative characteristic of useful financial reports is their ability to faithfully represent the underlying economic activities of the firm (e.g. IASB 2018, para. 2.12; FASB 2019, CON 8 Chapter 3). Therefore, I argue that accounting numbers can be used to identify differentiated firms because they should have *different* accounting numbers. In other words, the more (less) differentiated a firm is, the less (more) correlated its numbers should be with its benchmarking firms – industry peers. In this chapter, I develop a new methodology that uses fundamentals to identify differentiated firms. My approach is similar in spirit to Hoberg and Phillips (2016), who use product descriptions in firm 10-Ks to construct a product similarity measure. However, my approach is not exclusively restricted to *product* differentiation as I consider a wider variety of observable proxies. Thus, I aim to capture multiple sources (financing, investing or operating) of heterogeneity which can make firms appear dissimilar from their industry peers and reduce the ability to benchmark them.

I adopt a three-step methodology to generate a firm-year measure of differentiation. In the first step, for each industry and every year, I run a (logistic) model in order to identify *industry characteristics* – variables that are most useful to define this particular industry. In this step, I follow previous work on industry classifications (Krishnan and Press 2003; Kahle and Walkling 1996; Hrazdil et al. 2014; Guenther and Rosman 1994; Bhojraj et al. 2003; Schmalensee 1985; Clarke 1989) and use an array of financial statement ratios (return on assets, leverage, current ratio, asset turnover, SG&A intensity, cost of goods sold intensity, profit margin, size) and valuation multiples (price to earnings ratio, market to book ratio, enterprise value to sales) as potential industry characteristics. In contrast with previous studies on industry heterogeneity which typically used a univariate methodology (correlation or standard deviation) to determine which industry classification is more homogenous across accounting variables or to get contextualized peer groups (Ramnath 2002; Parrino 1997; Ecker et al. 2013; Carney and Young 2006; Albuquerque 2009), my approach offers the opportunity to measure intra-industry heterogeneity in a multivariate setting. This step reveals that, for example, firms in the "Oil & Gas" industry (*gics101020*) are typically characterized by capital intensity (higher than most other industries), current ratio (lower), sales volatility (higher) and SG&A intensity (lower). In the second step, I select significant variables (at 1% level) from the first-step logistic model and use these variables to determine the centroid of each GICS6 industry-year – what a prototypical "industry core" firm would look like. In the third and final step, I calculate the Euclidean distance between the centroid and a firm's position in a given year. The differentiation (*DIFF)* variable is the decile rank of the Euclidean distance and can be interpreted as the average extent to which the values of a firm's industry-specific ratios deviate from those of firms in the industry core.

3.2. Hypothesis development

3.2.1. Industry component of returns

The use of the industry component of returns has a long traditional history in finance theory to predict firms' stock returns (e.g. Roll 1988). Firms in the same industry have correlated returns and industry-specific economic events should be contemporaneously reflected in their price. In empirical applications (e.g. Durnev et al. 2004; Chun et al. 2008; Cohen and Lou 2012; Chen et al. 2016), industry returns proxy for industry-specific news, representing information shared among a group of peer firms within the same industry. Also, some studies provide more insights regarding information spillovers between peer firms. For example, Cohen and Frazzini (2008) show that investors can use the customer-supplier relationship between firms to predict returns, while Hameed et al. (2015) observe that "bellwether firms" forecast revisions have implications for peers' stock prices. Firms farther to the geographical cluster of their industry have lower levels of industry information in their price (Engelberg et al. 2018). In addition, Cohen and Lou (2012) show that relative to single-segment firms, the stock price of "complicated" (multiple-segments) firms adjusts with a lag to industry events, a finding they attribute to information processing costs and investors limited resources.

Differentiated firms represent firms that are farther from their industry cluster. They are firms heterogeneous to their industry peers. I argue that investors do recognize that these firms are different from the other firms in the industry. Thus, industry news should have smaller implications for their differentiated firms' stock prices. Also, I argue that differentiated firms represent a type of "complicated firms" through their intra-industry heterogeneity. Thus, the information processing costs for investors will be higher for differentiated firms. On the contrary, the relationship between industry news and industry core firms should be more straightforward resulting in less information processing costs for investors. Therefore, I predict that investors will need more time to assess the implications of industry news for differentiated firms. This should result in the incorporation of industry news realized in a less timely manner for differentiated firms than for industry core firms. I formulate the following hypotheses:

H3.1a: The contemporaneous association between industry news and firm-specific stock returns is smaller for differentiated firms than industry core firms.

H3.1b: The delayed association between industry news and firm-specific stock returns is larger for differentiated firms than industry core firms.

H3.1c: The total (contemporaneous plus delayed) association between industry news and firm-specific stock returns is smaller for differentiated firms than industry core firms.

3.2.2. Analysts

Recent publications from the academic literature and the business press highlight that industry knowledge is the primary trait for both buy-side and sell-side analysts (*Institutional Investors* annual surveys; Brown et al. 2015). More precisely, Brown et al. (2015) show that analysts keep specializing in industries as nearly half of the sell-side analysts cover only one industry. In fact, since analysts have less access to firm-level idiosyncratic information than the management team or (large) institutional investors, they tend to rely on industry peer firms to make forecasts (Ramnath 2002). Consequently, they develop an industry specialization to improve their ability to forecast earnings which makes them specialists of firms within industries. In a seminal paper, Clement (1999) shows that analysts forecast accuracy is negatively associated with the number of industries followed. Then, Boni and Womack (2006) underline that analysts are good to rank firms inside industries, reinforcing the importance of the link between firms inside industries. More recently, Bradley et al. (2017) emphasize the importance of work experience of analysts before being analysts. They show that analysts with prior work experience in the industry become better industry specialists.

In this chapter, I argue that firms that are more heterogeneous to their industry peers will be harder to analyze as analysts will not be able to rely on information from other firms to issue forecasts. According to the methodology developed in this chapter, firms inside industries have a unique position in a spatial representation of *n*-dimensions. Since analysts are industry specialists, I assume that they are more likely to be specialists of the typical firms in the industry. Consequently, firms farther from the industry core (i.e. differentiated firms) will have higher information processing costs since their financial reports are more difficult to be benchmarked against their industry peers. Due to this lack of benchmark, information provided by differentiated firms is less comparable cross-sectionally to their industry peers which makes them harder to analyze. I predict that differentiated firms will experience lower forecast accuracy and greater forecast dispersion than industry core firms. Also, I argue that covering differentiated firms will be costlier and riskier – since analysts' compensation is aligned with their forecasts accuracy (Brown et al. 2015) – which will reduce the overall coverage of differentiated firms compared to industry core firms.

H3.2a: Differentiated firms receive less coverage from analysts

H3.2b: Differentiated firms' forecasts are less accurate

H3.2c: Differentiated firms' forecasts face greater dispersion

3.2.3. Information asymmetry

Peterson et al. (2015) argue that a "*decreased homogeneity reduces the ability of market participants to make inferences from information disseminated from other firms in the industry, thereby resulting in greater information asymmetry"* (p. 2489). Consequently, in their cross-sectional tests they show that firms with a lower consistency compared to peer firms have a greater information asymmetry. Also, previous studies have shown that peer disclosures have spillover effects which reduce the information asymmetry between firms and investors for the relative peer group. For example, Shroff et al. (2017) show that peer information serves as a substitute for firms that have less publicly available firm-specific information (i.e. private firms in their setting). However, as the amount of firm disclosure increases – when the private firm becomes public – the need for peer information decreases.

I argue that these spillover effects are smaller for differentiated firms because their heterogenous nature may hinder market participants' ability to draw inferences from information disseminated by industry peers. In this context, peer information is unlikely to serve as a substitute for firm-specific information. Moreover, some firms may be differentiated for strategic purposes[20]. Thus, these firms experience higher proprietary costs to maintain their competitive advantage reducing incentives to voluntary disclose firm-specific information. Finally, since analysts serve as financial intermediaries between investors and insiders (Kelly and Ljungqvist 2012; Bradley et al. 2017), the low coverage of differentiated firms will increase the information asymmetry for market participants. Overall, I expect differentiated firms to experience a greater information asymmetry.

H3.3: Differentiated firms experience greater bid-ask spreads and have less liquidity

3.3. Research design

3.3.1. Construction of *DIFF*

I provide a new three-stage methodology to obtain a firm-year-specific differentiation score based on accounting and market fundamentals. A key assumption of this methodology is the

[20] Contrary to previous studies on differentiation (e.g. Hoberg and Philips, 2016) I do not restrict my proxy for differentiation to *product differentiation*. Thus, I argue that some firms may appear differentiated to their industry peers without voluntarily following a differentiation strategy.

multidimensional nature of industry heterogeneity. I assume that economic heterogeneity dimensions not only come from financial statement ratios but could also contain financial information or more strategy-related information. A multidimensional construct provides a framework that could potentially capture information that is relevant to peer selection and obtain an objective measure of economic heterogeneity. More importantly, I posit that the dimensions used to measure heterogeneity are industry specific.

3.3.1.1. Stage 1: Industry Characteristics

As the first stage of my methodology, for every year I determine which variables are most representative of each GICS6 industry. To do so, I run the following logistic model yearly, for each GICS6 industry *j*:

$$\begin{aligned} IND(j)_i = {} & \beta_0 + \beta_1 Size_i + \beta_2 AT\_TURN_i + \beta_3 BtoM_i + \beta_4 CAP\_INT_i + \beta_5 COGS_i \\ & + \beta_6 CR_i + \beta_7 EP_i + \beta_8 EVS_i + \beta_9 LEV_i + \beta_{10} PM_i + \beta_{11} RD_i \\ & + \beta_{12} ROA_i + \beta_{13} \sigma(\text{CFO})_i + \beta_{14} \sigma(\text{SALES})_i + \beta_{14} XSGA_i + \varepsilon_i \end{aligned} \tag{3.1}$$

Where *IND(j)i* is a dummy variable equal to one if firm *i* belongs to the GICS6 industry *j*[21]. Therefore, every year, I estimate the model *j* times, always using the same (pooled) sample but alternating the dependent variable. I estimate the models on a yearly basis because industry membership may vary over time for a given firm and because GICS industry definitions can also change[22]. In this model, a positive (negative) coefficient on a variable implies that firms in a given industry typically have a higher (lower) value of this variable than firms in the overall population. I voluntarily choose the GICS6 classification as the peer selection method since previous literature reveals that it offers the best homogeneity and it is

[21] I use historical GICS data (at the 6-digit level) from Compustat to construct *IND(j)$_i$*.

[22] For example, in 2018 (effective from 1st of October) the GICS has experienced an important evolution. The industry group *gics2540 "Media"* and *gics451010 "Internet Software & Services"* has been discontinued and the industry group *gics5020 "Media & Entertainment"* (regrouping three GICS6 industries) has been created. See MSCI website for more details regarding the evolution of the GICS (https://www.msci.com/gics; section "Historical GICS Structure").

representative of the analysts' industry specialization[23] (Bhojraj et al. 2003; Boni and Womack 2006; Hrazdil and Scott 2013).

I aim to capture various dimensions of differentiation using accounting numbers. Therefore, the potential industry characteristics in the logistic model include both valuation multiples and financial statement ratios. First, as my measure aims to capture differences in business models and strategy, I use a large set of variables related to operating dimensions: asset turnover (*AT_TURN*); intangible intensity (*XSGA*); cost of good sold intensity (*COGS*); standard deviation of cash flows ($\sigma(CFO)$); standard deviation of sales ($\sigma(SALES)$); capital intensity (*CAP_INT*); research and development expenses (*RD*). Also, as average profitability varies across industries (Guenther and Rosman 1994; Bhojraj et al. 2003; Krishnan and Press 2003; Hrazdil et al. 2014; Clarke 1989; Hoberg and Phillips 2016), I use several ratios to capture profitability : return on assets (*ROA*); profit margin (*PM*); and enterprise value-to-sales (*EVS*). Similarly, as financial structure can be seen a source of differentiation as I expect homogeneous groups of firms to possess similar financial structure (Guenther and Rosman 1994; Kahle and Walkling 1996; Bhojraj et al. 2003; Hrazdil et al. 2014; Krishnan and Press 2003), I use the following ratios to capture these differences: current ratio (*CR*); and leverage (*LEV*). Also, I add the following valuation ratios as they can be used in peer selection (Bhojraj and Lee 2002): book-to-market (*BtoM*); and earnings-to-price ratio (*EP*). Finally, Size is added because it has been used in previous studies as both a main and secondary criteria to form homogeneous groups of firms (Albuquerque 2009; Ecker et al. 2013).

3.3.1.2. Stage 2: Industry Centroid

As the second stage of my methodology, for each industry-year I select only significant variables (at the 1% level[24]) from the stage 1 logistic model as input variables (i.e. industry

[23] As a robustness test, I use the SIC2 as a starting point to create *DIFF*. In untabulated results, I observe that the two *DIFF* measures correlate at 0.44 while all the regression results (from Section 3.4.2) remain qualitatively similar.

[24] Some variables exhibit an important correlation that could create multicollinearity problems. This is particularly worrying since I choose a 1% threshold to select the significant variables. Thus, it could lead to the exclusion of variables that should have been included as a ratio characterizing the industry. In untabulated results, I compute the variable *DIFF* using a 5% and 10% threshold on the p-value and show that they correlate with the initial *DIFF* (i.e. using a 1% threshold) at 0.83 and 0.77, respectively.

characteristics) to determine the industry centroid. This centroid is basically defined as a hypothetical point minimizing the distance between each firms on a spatial representation of firms where the dimensions are the input variables. I interpret this centroid as the "prototype" firm of each industry-year – the "best" representative (from a statistical standpoint) of this industry-year's core. I make the strong assumption that only one centroid – where its spatial coordinates are the means of each dimension – exists for each industry-year. This assumption is debatable since GICS at the 6-digit level – or similarly the two-digit SIC – may suffer from within-industry homogeneity which could be interpreted as the existence of several centroids (see for example Bhojraj et al. 2003; Owens et al. 2017). However, this level of industry seems to be aligned with analysts' practice (Boni and Womack 2006; Kadan et al. 2012) and there is no evidence that other market participants use lower levels of industries – smaller peer groups.

3.3.1.3. Stage 3: Differentiation (distance) Score

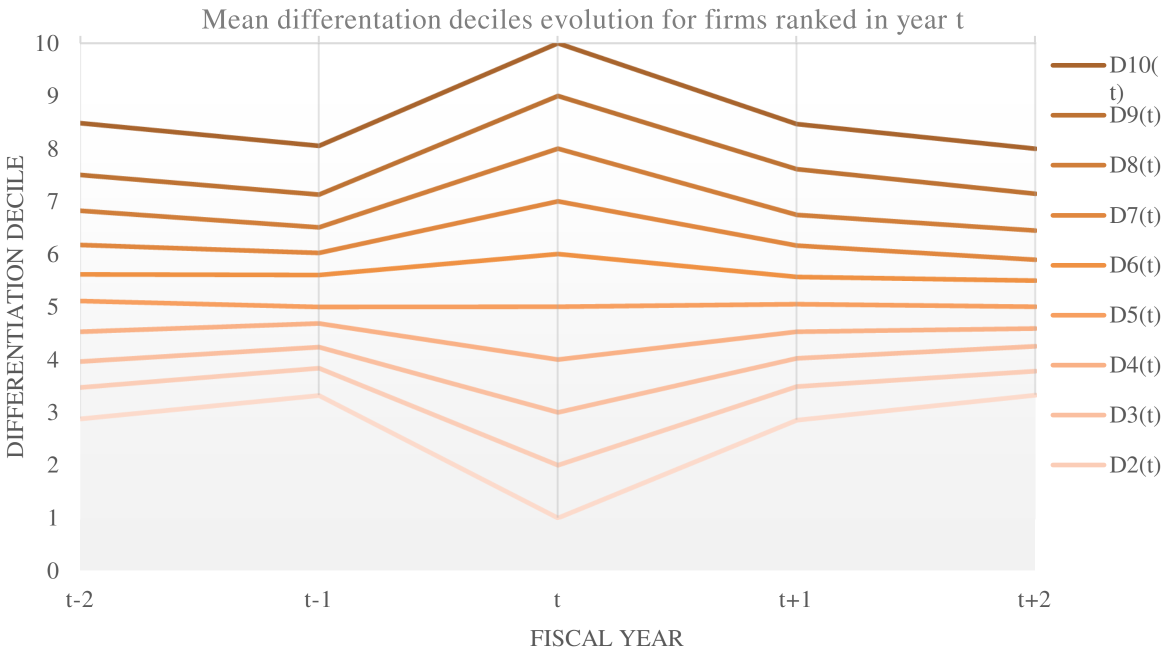
Finally, the third stage of my methodology consists of measuring how far other firms are from the stage 2 "prototype" firm. I use Euclidean distance, which represents the distance between two points in a Euclidean n-space. The Euclidean distance, in year t, between firm L's position in the n-space and its industry centroid M is computed as follows:

$$d(L_t, M_t) = \sqrt{\sum_{v=1}^{n} (L_{iv} - M_v)^2} \tag{3.2}$$

Where *v* indexes dimensions in the n-space (i.e. each *v* is a ratio characterizing an industry).

In this context, the distance represents how different a firm is from its (GICS6) industry peers. In other words, the farther (closer) a firm is from the "prototype" firm, the more (less) differentiated it is. Ultimately, I rank this distance into deciles to obtain *DIFF*, where the first (tenth) decile D1 (D10) represents the closest (farthest) firms to the "prototype". D1 (*DIFF*=1*)* and D10 (*DIFF*=10) firms are later referred to as industry core firms and differentiated firms, respectively. Even if *DIFF* is firm-year-specific, as we can see through Fig 3.1 it is stable through time as it is highly correlated with its lagged value[25]. Thus, *DIFF* helps to identify long-term differentiated firms rather than operational shocks.

Figure 3.1: Evolution of mean differentiation deciles for firms ranked in year t



This figure presents the evolution of differentiation deciles through time. First, firms are ranked into deciles at year t using their differentiation distance. Then, for each deciles the mean ranked value is calculated up to two years before and after (from t-2 to t+2).

3.3.2. Empirical models

3.3.2.1. Industry components of returns

In order to test H3.1a, H3.1b and H3.1c, I follow Cohen and Lou's (2012) methodology and estimate the following model:

[25] The (untabulated) correlation between the DIFF and its lagged value is 0.56.

$$RET_{i,t} = \beta_0 + \beta_1 RET_{i,t-1} + \beta_2 INDRET_{i,t} + \beta_3 INDRET_{i,t-1} + \beta_4(DIFF_{i,t} * INDRET_{i,t}) + \beta_5(DIFF_{i,t} * INDRET_{i,t-1}) + \beta_6(DIFF_{i,t} * RET_{i,t-1}) + \beta_7 DIFF_{it} + \beta_8 Size_{it} + \beta_9 BtoM_{it} + \beta_{10} Turnover_{it} + \varepsilon_{it} \tag{3.3}$$

$RET_{i,t}$ represents the return of firm *i* for day *t,* and $IND\_RET_{i,t}$ ($IND\_RET_{i,t-1}$) is the value-weighted return for firm *i*'s industry on day *t* (*t*-1)[26]. All the other variable definitions are in Appendix A. Standard errors are clustered by firm and day. I also run a second model (Model (2) in Table 3.4) where $DIFF_{i,t}$ is replaced with its lagged value $DIFF_{i,t-1}$ – $DIFF$ from the previous year.

First, H3.1a predicts that the contemporaneous association between industry news and firm-specific stock returns is smaller for differentiated firms than industry core firms. The coefficient of interest is β_4 and should be negative to confirm the hypothesis. Moreover, I hypothesize that the delayed association between industry news and firm-specific stock returns is larger for differentiated firms than industry core firms. In other words, H3.1b should lead to a positive coefficient on the interaction term $(DIFF_{i,t} * IND\_RET_{i,t-1})$. Finally, H3.1c predicts that differentiated firms will have less industry news than industry core firms. Thus, in equation 3.3, the coefficients of interest are those on the interaction terms $(DIFF_{i,t} * IND\_RET_{i,t})$ and $(DIFF_{i,t} * IND\_RET_{i,t-1})$. To confirm hypothesis H3.1c, the sum of the coefficients β_4 and β_5 should be negative.

3.3.2.2. Analyst tests

To test H3.2a, H3.2b and H3.2c, I follow previous literature (De Franco et al. 2011 ; Peterson et al. 2015) and estimate the following models:

$$Coverage_t = \beta_0 + \beta_1 DIFF_t + \beta_2 Size_t + \beta_3 BtoM_t + \beta_4 Volume_t + \beta_5 R\&D_t + \beta_6 Depreciation_t + \beta_7 Issue\_3y_t + \beta_8 \sigma(ROA)_t + \beta_9 \sigma(RET)_t + \varepsilon_t \tag{3.4}$$

[26] I calculate a different industry returns for each firm-day, where the firm i returns is excluded from the industry returns.

Where *Coverage* represents the number of analysts covering the company *i* for year *t*. Control variables are defined in Appendix A.

$$\begin{aligned} Dep\_VAR_t = \ & \beta_0 + \beta_1 DIFF_t + \beta_2 Size_t + \beta_3 \sigma(ROA)_t + \beta_4 \sigma(RET)_t \\ & + \beta_5 SUE_t \\ & + \beta_6 NegSUE_t + \beta_7 LOSS_t + \beta_8 Issue\_3y_t + \beta_9 NegSI_t \\ & + \beta_{10} Days_t + \varepsilon_t \end{aligned} \tag{3.5}$$

Where Dep_VAR is either *Accuracy* or *Dispersion*[27] . Control variables are defined in Appendix A.

When *Dispersion* (*Accuracy* or *Coverage*) is the dependent variable, a positive (negative) coefficient on *DIFF* would be consistent with H3.2a, H3.2b and H3.2c. In addition to the model above, I also estimate alternative specifications. In the first, in order to reduce multicollinearity concerns between the contemporaneous *DIFF* variable and independent variables, I use its lagged value $DIFF_{t-1}$, i.e. firm *i*'s differentiation decile as constructed in the previous year. Finally, because my measure of differentiation aims to capture the *systematic* deviation of differentiated firms from their industry peers but may be affected by operational shocks experienced by a firm in a particular year, I introduce a final specification in which I replace $DIFF_t$ with its two components $DIFF_{t-1}$ and $\Delta DIFF_t$. To the extent that past differentiation is a stronger driver of analyst coverage, accuracy, or dispersion than current operational shocks, then the coefficient on $DIFF_{t-1}$ should be larger (in absolute terms) than that on $\Delta DIFF_t$.

3.3.2.3. *Information asymmetry tests*

In order to test the association between differentiation and information asymmetry, I follow previous literature (Peterson et al. 2015) and estimate the following model:

[27] I use the latest one year-ahead annual forecast available from each analyst before the earnings announcement to compute both *Accuracy* and *Dispersion*. Untabulated results suggest that these results are robust to the use of mean or median of every analyst forecasts issued during the fiscal year before the announcement date.

$$InfoAsym_t = \beta_0 + \beta_1 DIFF_t + \beta_2 MVE_t + \beta_3 \sigma(RET)_t + \beta_4 Turnover_{i,t} + \varepsilon_{it} \tag{3.6}$$

I use two different proxies for information asymmetry. First, I use Amihud's (2002) illiquidity measure averaged on a 12-month window starting two months before the fiscal year end. The second measure provided in this model is the bid-ask spread. As with tests on analysts, I run additional models where *DIFF* is replaced by its lagged value, or by its lagged value and a proxy for operational shocks ($\Delta DIFF_t$). As in Peterson et al. (2015) I include controls for market value of equity (*MVE*), returns volatility (*σ(RET)*) and share turnover (*Turnover*). Standard errors are clustered by firm and year.

3.3.3. Sample construction

I obtain accounting data from Compustat, market data from CRSP and analyst data from I/B/E/S. Starting from the initial sample used throughout this thesis[28], I also exclude observations with total assets fewer than 10 million, negative sales, market capitalization under 10 million, negative common stakeholder equity or with missing data to compute the differentiation score. Also, because the methodology I use to construct this score involves the estimation of logistic regressions for each industry, and model performance could be poor in industries with a small number of observations, I exclude observations in industry-years that contain less than 15 observations[29]. Ultimately, I am able to compute a value for *DIFF* for 22,165 firm-year observations[30]. Then, for each different test I delete missing observations from this specific test. Therefore, samples across subsequent tests (Table 3.4 to 3.6) can be different for each test.

[28] See section 2.2.3 for a description of the exclusion criteria.

[29] Despite this specification, in some years, some industries still have no significant variables in the first-stage model. I also exclude these observations since I am unable to compute a *DIFF* for them.

[30] An important proportion of firms (5,335 firm-year observations) is lost due to the computation of the lagged value of *DIFF*. I require this variable in subsequent tests to assess the robustness of my results.

TABLE 3.1: Sample selection

	Firm-year observations
Initial sample common to all chapters	98746
Less firms with negative sales; assets under 10 millions; market capitalization under 10 millions; or negative common equity	(44,518)
Less missing data for input variables of the logistic model	(22,472)
Less industries with less than 15 observations per year	(2,300)
Less industries with no significant variables in Step 1	(1,956)
Less missing observation for lagged value of differentiation (*DIFF*)	(5,335)
Full sample	22,165
Industry news incorporation tests sample	
Merge with CRSP daily returns	5,233,754[31]
Analyst tests sample	
Less missing data on *Accuracy* or *Coverage*	(4,567)
Less missing data on *Dispersion*	(2,506)
Full sample for analyst tests	15,092
Information asymmetry tests sample	
Less missing data on *Bid-Ask* or *Illiquidity*	(826)
Full sample for information asymmetry tests	21,339

3.4. Results

3.4.1. Descriptive statistics

Table 3.2 presents some descriptive statistics for the main variables[32]. On average, firms in the final sample exhibit a positive ROA (2,6%) and the mean (median) value for Size is 6.558 (6.499). Overall, means and medians are consistent with previous studies with similar sample restrictions. Also, this table shows that input variable used in the third stage to calculate the Euclidean distance exhibit high dispersion regarding their magnitude. Therefore, to avoid any impact of the magnitude I standardize each variable with a mean of 0 and a standard deviation of 1. This standardization aims to put the same weight on each industry characteristics and to calculate an equally weighted Euclidean distance across input variables.

[31] This represents firm-day observations.

[32] All the continuous variables are winsorized at 1% and 99%, except for *Accuracy* and *Dispersion*. For these variables, after a winsorization at 1% and 99% levels, the sample is still contaminated by outliers affecting the subsequent regression results. In order to have a distribution closer from previous studies (De Franco et al. 2011 ; Peterson et al. 2015), I apply a stricter threshold (5%-95%) for the winsorization of these variables.

TABLE 3.2: Descriptive statistics

Variable	N	Mean	StdDev	P10	Median	P90
DIFF	22,165	5.452	2.818	2.000	5.000	9.000
Size	22,165	6.558	1.844	4.137	6.499	9.021
AT_TURN	22,165	1.236	0.800	0.374	1.068	2.318
BtoM	22,165	0.617	0.542	0.178	0.478	1.170
CAP_INT	22,165	0.300	0.245	0.053	0.225	0.699
COGS	22,165	0.834	0.692	0.144	0.662	1.730
CR	22,165	1.828	1.164	0.630	1.596	3.236
EP	22,165	-0.019	0.294	-0.136	0.042	0.093
EVS	22,165	2.170	3.041	0.391	1.287	4.503
LEV	22,165	0.182	0.166	0.000	0.157	0.419
PM	22,165	-0.018	0.398	-0.112	0.038	0.144
RD	22,165	0.034	0.120	0.000	0.000	0.081
ROA	22,165	0.026	0.123	-0.092	0.046	0.129
σ(CFO)	22,165	0.050	0.044	0.012	0.037	0.104
σ(SALES)	22,165	0.142	0.141	0.028	0.097	0.305
XSGA	22,165	0.278	0.234	0.041	0.215	0.603
Accuracy	17,821	-1.253	3.793	-2.451	-0.194	-0.010
Coverage	17,821	1.679	0.972	0.000	1.792	2.890
Dispersion	15,349	0.365	0.828	0.020	0.106	0.775
BidAsk	21,316	0.004	0.006	0.000	0.001	0.011
Illiquidity	21,316	0.054	0.142	0.000	0.003	0.157

This table presents descriptive statistics for industry characteristics and dependent variable from the regression results. All variables are winsorized at 1% and 99%, except for Accuracy and Dispersion winsorized at 5% and 95%.

3.4.2. Methodology results

To gain an understanding of the results of stage 1 logistic regressions (Eq. 3.1) and determine whether industry characteristics are stable over time, I calculate the statistics summarized in Table 3.3. For each industry, I calculate the percentage of years (out of 19 possible years from 1999 to 2018) for which each variable is significantly associated with industry membership; for brevity, I only present the three industries where each variable is the most frequently significant, for both positive and negative coefficients[33] [34]. For example, the coefficient on capital intensity (*CAP_INT*) is positive (negative) and significant for all years (100%) for the *gics101020 Oil, Gas & Consumable Fuels (gics254010 Media)* industry,

[33] Some industries do not have any significant variables. Three explanations are plausible for that. First, in order to obtain the most precise measure of differentiation, I use a very constraining level of significance (p-value <1%). Second, these industries have a small number of observations, which could lead to a lack of power in models' performances. Third, these industries may possess financial ratios that do not differ on any dimensions from the rest of the firms. Ultimately, for years when industries do not have any significant variable, I delete these observations from the final sample, as I cannot compute the Euclidean distance. In order to avoid bias from this design choice, I run additional (untabulated) tests integrating these observations into the mid-decile (D6). Results remain qualitatively similar.

[34] Four industries are reported when multiple industries are tied for 3rd.

indicating that firms in the oil and gas (media) sector typically have a higher (lower) proportion of fixed assets than firms in the overall population.

Through Table 3.3, several patterns can be observed. For example, *CR* is always positively or negatively significant in some industries, suggesting large differences in the average current ratio across industries. Capital intensity (*CAP_INT*), current ratio (*CR*), enterprise value-to-sales (*EVS*) and research and development expenses (*RD*) represent the four most frequently significant variables. Thus, these variables seem to play a systematic role in explaining industry membership. On the contrary, the standard deviation of cash flows (*σ(CFO)*) or sales (*σ(SALES)*), or the earnings-to-price ratio (*EP*) are rarely significant. However, every variable has both positive and negative significant coefficient in at least one industry-year, suggesting that they all play a role – even if minor – in characterizing industry membership.

Naturally, research and development expenses (*RD*) is positively associated for every year with industries *gics 352010 "Biotechnology"* and *gics 201010 "Aerospace & Defence"*, and is regularly negatively associated with retail and services industries (e.g., *gics 255030 "Multiline Retail"; gics 451020 "IT Services"*). Also, results on *XSGA* are consistent with previous literature (Srivastava 2014; Enache and Srivastava 2018), as the *"Food & Staples Retailing" industry (gics 301010)* – comparable to Fama-French "Food Products" industry – is positively associated with *XSGA* in 76% of the yearly regressions, while the *"Oil, Gas & Consumable Fuels"* industry *(gics 101020)* – comparable to the Fama-French "Petroleum and natural gas" industry – is negatively associated with *XSGA* in every yearly regression. This result is consistent with the intuition that firms in the retail industry have higher selling, general and administrative expenses than in the petroleum industry. Overall, significant variables are consistent with the intuition behind each industry. Other variables, such as *ROA*, are more rarely significant in Eq (3.1), which either indicates a large dispersion of within-industry profitability or a distribution close to the full sample[35]. More precisely, it points out

[35] I acknowledge that some insignificant coefficients could come out of a high correlation between some variables, which may inflate standard errors in the logistic regressions. However, in an untabulated table, I find that only 5 pairs of variables exhibit an absolute correlation higher than 0.5 (*PM* with *ROA, EVS* and *RD*; *ROA* with *EP*; *AT_TURN* with *COGS*). Overall, it should have a minor impact on the *DIFF* variable. In untabulated results, I alternatively exclude *AT_TURN, COGS, ROA* and *PM* variables. Results from the regressions remain qualitatively similar.

that no industries outperform (or underperform) the market systematically over the sample period. At any rate, because an insignificant stage 1 coefficient implies that the variable is disregarded in the subsequent stages and in the construction of *DIFF*, this means that firms with extreme *ROA* values are generally not considered as “differentiated” firms for that reason alone.

TABLE 3.3: Industry characteristics: Frequencies of significant coefficients in industry-year regressions

Variable	Overall significance	Top 3 industries		
Size	-	351030	Health Care Technology	71%
		202010	Commercial Services & Supplies	95%
	(10%)	253010	Hotels, Restaurants & Leisure	100%
	+	301010	Food & Staples Retailing	53%
		255040	Specialty Retail	94%
	(12%)	255030	Multiline Retail	100%
AT_TURN	-	302010	Beverages	12%
		201040	Electrical Equipment	29%
	(5%)	201030	Construction & Engineering	35%
	+	351020	Health Care Providers & Services	17%
		101020	Oil, Gas & Consumable Fuels	18%
	(3%)	303020	Personal Products	18%
BtoM	-	151030	Containers & Packaging	47%
		201060	Machinery	50%
	(17%)	351010	Health Care Equipment & Supplies	78%
	+	253010	Hotels, Restaurants & Leisure	6%
	(1%)	255030	Multiline Retail	29%
CAP_INT		254010	Media	100%
	-	201060	Machinery	100%
		252030	Textiles, Apparel & Luxury Goods	100%
	(38%)	351010	Health Care Equipment & Supplies	100%
	+	253010	Hotels, Restaurants & Leisure	78%
		301010	Food & Staples Retailing	88%
	(14%)	101020	Oil, Gas & Consumable Fuels	100%
COGS	-	352010	Biotechnology	29%
		352020	Pharmaceuticals	47%
	(10%)	303020	Personal Products	82%
	+	255020	Internet & Direct Marketing Reta	24%
		201030	Construction & Engineering	29%
	(4%)	201040	Electrical Equipment	29%
CR	-	351020	Health Care Providers & Services	78%
		254010	Media	88%
	(17%)	253010	Hotels, Restaurants & Leisure	100%
	+	252030	Textiles, Apparel & Luxury Goods	94%
		351010	Health Care Equipment & Supplies	94%
	(31%)	151040	Metals & Mining	100%
EP	-	252020	Leisure Products	24%
		352010	Biotechnology	24%
	(6%)	255040	Specialty Retail	33%
	+	151050	Paper & Forest Products	14%
		251010	Auto Components	12%
	(2%)	101020	Oil, Gas & Consumable Fuels	24%
EVS	-	253010	Hotels, Restaurants & Leisure	89%
		151030	Containers & Packaging	94%
	(31%)	201060	Machinery	100%
	+	254010	Media	12%
		255020	Internet & Direct Marketing Reta	12%
	(2%)	101020	Oil, Gas & Consumable Fuels	41%

TABLE 3.3 (continued)

Variable	Overall significance	Top 3 industries		
LEV	-	151040	Metals & Mining	35%
		255040	Specialty Retail	61%
	(8%)	351010	Health Care Equipment & Supplies	67%
	+	151030	Containers & Packaging	35%
		501010	Diversified Telecommunication Se	41%
	(5%)	351020	Health Care Providers & Services	44%
PM	-	301010	Food & Staples Retailing	12%
		302020	Food Products	16%
	(5%)	255030	Multiline Retail	57%
	+	201010	Aerospace & Defence	17%
		351010	Health Care Equipment & Supplies	33%
	(6%)	352010	Biotechnology	47%
RD	-	255030	Multiline Retail	86%
		202010	Commercial Services & Supplies	68%
	(17%)	252030	Textiles, Apparel & Luxury Goods	83%
	+	352020	Pharmaceuticals	41%
		201010	Aerospace & Defence	78%
	(12%)	352010	Biotechnology	94%
ROA	-	101020	Oil, Gas & Consumable Fuels	18%
		251010	Auto Components	18%
	(4%)	101010	Energy Equipment & Services	24%
	+	255030	Multiline Retail	29%
		255040	Specialty Retail	28%
	(7%)	201060	Machinery	33%
σ(CFO)	-	351010	Health Care Equipment & Supplies	22%
		201060	Machinery	28%
	(4%)	253010	Hotels, Restaurants & Leisure	28%
	+	252030	Textiles, Apparel & Luxury Goods	22%
		151040	Metals & Mining	24%
	(3%)	352020	Pharmaceuticals	35%
σ(SALES)	-	253010	Hotels, Restaurants & Leisure	22%
		151030	Containers & Packaging	24%
	(6%)	301010	Food & Staples Retailing	24%
	+	201060	Machinery	11%
		151040	Metals & Mining	18%
	(5%)	101020	Oil, Gas & Consumable Fuels	100%
XSGA	-	251010	Auto Components	47%
		101010	Energy Equipment & Services	100%
	(14%)	101020	Oil, Gas & Consumable Fuels	100%
	+	255020	Internet & Direct Marketing Reta	71%
		301010	Food & Staples Retailing	76%
	(12%)	255040	Specialty Retail	78%

This table reports the frequencies of positive and negative significant coefficients of the logistic regression for each variable. Coefficients are considered significant for years where the *p-value* is below 1%. In column "Top 3 industries" I present only the three industries where the corresponding variable is the most frequently significant, for both negative and positive associations. In column "Overall significance", frequencies between parentheses below +(-) represent the positive (negative) significant coefficients frequencies for each variable across every industry year regressions.

Appendix C provides an example for the calculation of *DIFF* for the industry *gics 251010 "Auto Components"* for fiscal year 2006. Panel A presents the results of the logistic regression of the first stage, where we can see that only two variables (*RD and XSGA*) are significant at 1% level. Consequently, these two variables represent the ratios characterizing the membership to this industry in 2006. Panel B provides descriptive statistics for *RD* and *XSGA* to highlight the differences between the full sample and the industry, and across the differentiation deciles. Finally, Panel C reports a two-dimensions graph – where *RD* (*XSGA*) is on the x-axis (y-axis)[36] – illustrating the spatial distributions of firms in this industry, where gray (black) dots represent differentiated (industry core) firms. Through this graph, we see clearly that differentiated firms in this industry represents firms having either high deviation of *RD* or *XSGA*, or both. For example, differentiated firms (Q5) have either very high value of *XSGA* or *RD*. This suggests that they differentiate from their intra-industry peers through a higher level of intangible assets, or through a higher level of research and development expenses.

3.4.3. Regression results

3.4.3.1. Industry component of returns

Table 3.4 presents the results of a regression of daily returns on industry returns, and control variables. This model aims to evaluate the incorporation of industry news into differentiated firms' stock prices. The coefficients on contemporaneous and lagged industry news are both positive and significant, indicating that daily stock returns of firms in the industry core are affected by today's, and to some extent yesterday's, industry news. Coefficient β_4 on variable $INDRET_t*DIFF_t$ is negative (-0.0152) and significant at 1% level, indicating that differentiated firms possess less contemporaneous industry news than industry core firms. This result confirms H3.1a. The positive and significant (at 1%) coefficient β_4 (0.0046) on $INDRET_{t-1}*DIFF_t$ means that today's price is more affected by yesterday's industry news for differentiated firms than for industry core firms. This result is in accordance with H3.1b. Then, the sum of coefficients β_4 and β_5 (-0.0106) is negative and confirms that overall, industry news play a smaller role in explaining returns for differentiated firms than for

[36] The centroid position for each industry-year is the mean of each variable. In this graph, I present the spatial coordinates of each firm with their standardized value. Thus, the centroid position is (0,0).

industry core firms. This result confirms H3.1c. Also, the significant coefficient β_5 has other implications for the literature. This highlights that using only the contemporaneous value of industry returns is not sufficient to fully capture the comovement of a firm with its industry since using only the contemporaneous value (β_4) of industry returns underestimates the overall impact ($\beta_4 + \beta_5$) of industry news on differentiated firms. Consequently, without including the lagged value of industry news, differentiated firms are presented as less related to their industry peers than they really are. Results hold when $DIFF_{t-1}$ is used instead of $DIFF_t$ as in model (2).

TABLE 3.4: Industry news incorporation

Model		**(1)**			**(2)**	
Variable	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
RET_{t-1}	-0.0367	***	-9.0	-0.0365	***	-9.3
$INDRET_t$	0.9755	***	69.5	0.9627	***	68.9
$INDRET_{t-1}$	0.0380	***	5.9	0.0395	***	6.2
$INDRET_t*DIFF_t$	**-0.0152**	***	**-7.2**			
$INDRET_{t-1}*DIFF_t$	**0.0046**	***	**6.0**			
$INDRET_t*DIFF_{t-1}$				**-0.0128**	***	**-6.2**
$INDRET_{t-1}*DIFF_{t-1}$				**0.0043**	***	**5.4**
$RET_{t-1}*X$	-0.0014	***	-2.8	-0.0015	***	-2.9
X	-0.0000		-1.5	0.0000		1.3
Size	-0.0056	*	-1.8	-0.0060	*	-1.9
BtoM	-0.0433	***	-12.3	-0.0441	***	-12.6
Turnover	0.0058		1.2	0.0062		1.3
Intercept	-0.0001		-1.3	-0.0002	**	-2.5
Adjusted R^2	0.22			0.22		
N	5,233,754			5,233,754		

Table 3.4 reports the results regarding the industry component of returns. Model (1) uses the contemporaneous value of *DIFF*. Model (2) provides a robustness test using the lagged value of *DIFF*. *Size*, *BtoM* and *Turnover* variables are scaled by 100 to get more understandable coefficients in this regressions since daily returns are used. Standard errors are double-clustered by firm and time. *, **, *** represent significance at 1%, 5% and 10%, respectively.

3.4.3.2. Analysts

Table 3.5 provides evidence regarding the information processing from analysts. When *Coverage* is the dependent variable (Panel A), the coefficient on *DIFF* or its lagged value is negative and significant at 1% for each model. These results suggest that differentiated firms receive less coverage from analysts than industry core firms. Also, in model (3), the coefficients on $DIFF_{t-1}$ and $\Delta DIFF_t$ are both negative and significant. Then, it suggests that the lower coverage experienced by differentiated firms is driven by both its long-term differentiation and operational shocks. When *Accuracy* is the dependent variable (Panel B), the coefficients on both contemporaneous and lagged values of *DIFF* are once again negative and significant at the 1% level), even after controlling for operational shocks in model (3). These results suggest that analysts forecast earnings of differentiated firms in a less accurate fashion than industry core firms. This is consistent with the argument that the lack of appropriate benchmarks makes differentiated firms more difficult to predict. The coefficient on $\Delta DIFF_t$ is also negative and significant (at 1% level), confirming that firms experiencing operational shocks have earnings that are more difficult to forecast. Finally, when *Dispersion* is the dependent variable (Panel C), differentiated firms are positively associated (at 1%) with a greater dispersion in analysts' forecasts, a finding that is once again more strongly associated with the lagged (i.e. systematic) component of *DIFF* than with its current-year variation (i.e. operational shocks). Other control variables are in accordance with previous literature.

TABLE 3.5: Analyst tests

Panel A: Coverage

Model	**(1)**			**(2)**			**(3)**		
Variable	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
$\boldsymbol{DIFF_t}$	**-0.0089**	***	**-3.1**						
$\boldsymbol{DIFF_{t-1}}$				**-0.0101**	***	**-3.7**	**-0.0120**	***	**-3.5**
$\boldsymbol{\Delta DIFF_t}$							**-0.0045**	**	**-2.0**
Size	0.0074		0.4	0.0067		0.3	0.0086		0.4
BtoM	-0.1146	***	-6.7	-0.1179	***	-6.9	-0.1164	***	-6.8
Volume	0.4215	***	47.5	0.4215	***	47.6	0.4213	***	47.4
RD	0.0042		0.3	0.0030		0.2	0.0055		0.4
Depreciation	1.5267	***	4.1	1.5302	***	4.1	1.5208	***	4.1
Issue_3y	-0.0224		-0.8	-0.0227		-0.9	-0.0226		-0.9
$\sigma(ROA)$	-1.7360	***	-9.5	-1.7318	***	-9.4	-1.7216	***	-9.3
$\sigma(RET)$	-0.4248	***	-15.5	-0.4252	***	-15.7	-0.4242	***	-15.7
Intercept	-1.6724	***	-14.0	-1.6674	***	-14.3	-1.6535	***	-14.0
Adjusted R^2	0.61			0.61			0.61		
N	17,598			17,598			17,598		

Panel B: Forecast accuracy

Model	(1)			(2)			(3)		
Variable	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
$DIFF_t$	**-0.0606**	***	**-4.0**						
$DIFF_{t-1}$				**-0.0525**	***	**-3.9**	**-0.0710**	***	**-4.2**
$\Delta DIFF_t$							**-0.0454**	***	**-2.9**
Size	0.2180	***	2.9	0.2038	***	2.7	0.2212	***	2.9
σ(ROA)	-5.5190	***	-3.4	-5.5514	***	-3.4	-5.4671	***	-3.4
σ(RET)	-1.4067	***	-6.9	-1.4137	***	-6.9	-1.4057	***	-6.9
SUE	-0.0214	***	-3.3	-0.0211	***	-3.2	-0.0213	***	-3.3
NegSUE	0.0620		0.9	0.0587		0.9	0.0598		0.9
Loss	-1.4826	***	-8.1	-1.4887	***	-8.2	-1.4780	***	-8.1
NegSI	-3.7140		-1.5	-3.8122		-1.5	-3.7769		-1.5
Days	-0.3072	***	-6.2	-0.3061	***	-6.2	-0.3058	***	-6.2
Intercept	-4.4971	***	-5.5	-4.5615	***	-5.6	-4.4438	***	-5.5
Adjusted R^2	0.18			0.18			0.18		
N	17,598			17,598			17,598		

Panel C: Forecast dispersion

Model	(1)			(2)			(3)		
Variable	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
$DIFF_t$	**0.0122**	***	**3.6**						
$DIFF_{t-1}$				**0.0117**	***	**3.3**	**0.0149**	***	**3.5**
$\Delta DIFF_t$							**0.0081**	***	**3.4**
Size	-0.0592	***	-2.6	-0.0566	**	-2.6	-0.0604	***	-2.7
σ(ROA)	1.5502	***	4.4	1.5505	***	4.4	1.5357	***	4.4
σ(RET)	0.3099	***	5.2	0.3108	***	5.3	0.3097	***	5.2
SUE	0.0066	***	7.7	0.0065	***	7.7	0.0066	***	7.8
NegSUE	-0.0173		-0.9	-0.0169		-0.9	-0.0169		-0.9
Loss	0.5091	***	9.9	0.5099	***	9.8	0.5080	***	9.9
NegSI	-1.5036	***	-4.0	-1.4778	***	-4.0	-1.4857	***	-4.0
Days	0.0193	**	2.0	0.0190	**	2.0	0.0190	**	2.0
Intercept	1.2286	***	4.9	1.2342	***	5.0	1.2146	***	4.9
Adjusted R^2	0.20			0.20			0.21		
N	15,092			15,092			15,092		

Table 3.5 reports the results regarding the analyst forecasts accuracy and dispersion. Panel A presents the results when using *Coverage* as a dependent variable. Panel B presents the results when using *Accuracy* as a dependent variable. Panel C presents the results when using *Dispersion* as a dependent variable. For each panel, Model (1) uses the contemporaneous value of *DIFF*, Model (2) provides test of robustness using the lagged value of *DIFF*, and Model (3) uses both the lagged value of *DIFF* and an additional control for operational shocks. Standard errors are double-clustered by firm and time. *, **, *** represent significance at 1%, 5% and 10%, respectively.

3.4.3.3. Information Asymmetry

Table 3.6 presents the results regarding the information asymmetry tests. Panel A presents the estimation with *Illiquidity* as a dependent variable. I observe a positive and significant association (at 1% level) between *DIFF* or its lagged value, and *Illiquidity*. For model (3), the coefficient on $\Delta DIFF$ is positive and significant, which means that operational shocks increase *Illiquidity*. Panel B relates to the use of *BidAsk* as a dependent variable. Similarly, a significant positive association (at 1% level) is observed with my variables of interest (*DIFF* or its lagged value). Again, operational shocks are positively associated with *BidAsk* (model (3)). For both panels, the coefficients on control variables are consistent with prior literature. Overall, these results support hypothesis H3.3, and suggest that differentiated firms experience a greater illiquidity and a larger bid-ask spread. Therefore, they suffer from a greater information asymmetry.

TABLE 3.6: Information asymmetry test

Panel A: Illiquidity

Model	**(1)**			**(2)**			**(3)**		
Variable	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
$DIFF_t$	**0.0025**	***	**5.6**						
$DIFF_{t-1}$				**0.0023**	***	**5.0**	**0.0030**	***	**5.4**
$\Delta DIFF_t$							**0.0018**	***	**4.8**
MVE	-0.0096	***	-7.2	-0.0097	***	-7.3	-0.0096	***	-7.2
σ(RET)	0.0714	***	17.9	0.0717	***	17.8	0.0713	***	17.9
Turnover			-			-			-
	-0.0792	***	24.8	-0.0793	***	24.9	-0.0791	***	24.8
Intercept	0.5566	***	31.4	0.5595	***	30.8	0.5529	***	31.1
Adjusted R2	0.50			0.50			0.50		
N	21,339			21,339			21,339		

Panel B: Bid-Ask spread

Model	**(1)**			**(2)**			**(3)**		
Variable	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
$DIFF_t$	**0.0114**	***	**7.2**						
$DIFF_{t-1}$				**0.0099**	***	**6.5**	**0.0133**	***	**7.0**
$\Delta DIFF_t$							**0.0085**	***	**6.4**
MVE	-0.0556	***	-8.3	-0.0558	***	-8.4	-0.0555	***	-8.3
$\sigma(RET)$	0.3973	***	15.0	0.3989	***	15.1	0.3967	***	15.0
Turnover			-			-			-
	-0.3513	***	19.1	-0.3516	***	19.2	-0.3509	***	19.1
Intercept	3.0703	***	25.1	3.0863	***	25.3	3.0554	***	24.9
Adjusted R2	0.61			0.61			0.61		
N	21,339			21,339			21,339		

Table 6 reports the results of the information asymmetry tests. Panel A presents the results using the Illiquidity measure from Amihud (2002) as dependent variable. Panel B presents the results using the Bid-Ask spread as dependent variable. Dependent variables are multiplied by 100 to get more understandable regressions coefficient. Standard errors are double-clustered by firm and time. *, **, *** represent significance at 1%, 5% and 10%, respectively.

3.5. Summary

First, using a logistic model I present an industry-specific list of ratios characterizing the membership to this particular industry. Using this list, for each industry year I identify firms which are the most heterogeneous to their industry which I refer to differentiated firms. Through this new methodology, I provide new evidence regarding industry classifications. Some studies (e.g. Albuquerque (2009)) improve the intra-industry homogeneity through the similarly use of one variable (Size) to *every* industry. By showing that industries have specific characteristics of membership, I provide evidence that the use of industry-year specific variables can enhance intra-industry homogeneity resulting in more homogenous peer groups.

Second, I investigate how differentiation affects the information processing from market participants and I provide evidence regarding the economic consequences of being differentiated. I show that differentiated firms have harder information to process, which create an informational cost. More precisely, I show that analyst forecasts are less accurate and more disperse for differentiated firms, and I show that analysts are less willing to cover these firms. In a similar spirit, I show that differentiated firms have a greater information asymmetry which should result in a higher cost of capital.

Due to its exploratory nature, my study is subject to several limitations which also open multiple perspectives for future research. First, despite adding controls for operational shocks, I cannot totally rule out that some part of my differentiation measure is driven by operational performance in a given year. Nevertheless, the evidence suggests that my measure captures *systematic* deviations from industry standards, given the strong correlation between my measure and its lagged value and the similar results obtained when this lagged value is used. Secondly, Euclidean distance gives a simple way to construct a multivariate measure of homogeneity, but its main drawback is that it does not put any weight on each variable – each variable has the same weight in the final measure of distance. For example, one could think that research and development expenses is a *more important* determinant of an industry than leverage.

4. Conglomerates complexity

4.1. Introduction

In contrast with studies of linguistic and accounting complexity, in which the generation and evaluation of new proxies for complexity is often a central part of the investigation (e.g., Peterson 2012; Loughran and McDonald 2014; Filzen and Peterson 2015; Hoitash and Hoitash 2018), studies on operational complexity generally rely on simple proxies that are widely available, such as the number of business segments (*NBS*), firm size or the existence of foreign operations (see for example Hoitash and Hoitash 2018 - Online Appendix 1 for an extensive literature review of research on complexity). Thus, although operational complexity is a determinant of financial reporting complexity (Guay et al. 2016), it surprisingly appears as an understudied field, and the use of oversimplified proxies may lead to some of the insignificant results documented in the literature.[37] Additional research is even more appealing since financial statement complexity seems to be more related to the innate (operational) complexity of the firm than a desire by managers to voluntarily obfuscate information (Bushee et al. 2018).

In this chapter, I provide a new measure that challenges the fundamental assumptions underlying the use of *NBS* (or the natural logarithm of *NBS*) as a proxy for operational complexity. Counting the number of business segments does not account for the *extent* to which segments differ from each other: firms with the same *NBS* are deemed equally complex even if one firm has two segments in very similar industries while another has segments in two very different industries. In this chapter, I posit that the operational complexity of multiple-segment firms depends on the relatedness of the industries in which the firm's individual segments operate.

My new measure, *Log_CPX*, is a firm-year measure of the natural logarithm of *the number of complex (secondary) business segments*. To obtain *Log_CPX*, I first calculate the distance

[37] Hoitash and Hoitash (2018) review papers published in leading accounting journals from 2004 to 2014 and find that operating complexity measures are rarely included in regression models of financial reporting quality, and that when they are included in such models, their coefficients are rarely significant. Under the assumption that operating complexity has a *potentially* detrimental effect on quality, this either implies that preparers anticipate this potential effect and allocate the appropriate level of resources to eliminate it, that operating complexity is measured with error, or both.

between each pair of SIC2 industries. This distance is based on the relative locations of both industries in an *n*-dimensional space, where locations' coordinates are given by the mean values of *n* fundamental ratios for each industry's single-segment firms. The chosen ratios characterize industries on the operating, financial, profitability and valuation dimensions. A larger (smaller) distance indicates that, on average, the ratios of the two industries are farther from (closer to) each other, suggesting that both industries are less (more) likely to be related. Finally, I classify as *complex* all segments for which the distance between the business segment industry and the firm's primary industry is higher than the median distance between every possible pairs of SIC2 industries.

4.2. Hypothesis development

Despite being seen as a determinant of financial statement complexity, operational complexity remains an understudied field (Guay et al. 2016 ; Hoitash and Hoitash 2018). As argued by Hoitash and Hoitash (2018), "because detailed disclosures of firm operations are not widely available, researchers often rely on observable measures of operating complexity" (p. 263), such as the number of business or geographic segments. With SFAS 131 on segment reporting, the FASB adopted a management approach stating that segment reporting should be aligned with how top management evaluates the performance of business units within the firms. Thus, the number of business segments reported appears as a natural proxy for operational complexity since it should reflect the complexity of the internal organization of the firm – see for example in the context of auditors (Choi et al. 2010; Hoitash and Hoitash 2018) or analysts (e.g. Lehavy et al. 2011; Dechow and You 2012). This association between the number of business segments and complexity suggests that conglomerates – firms that have more than one business segment – are complex firms. However, recent literature challenges the assumption that conglomerates operate in business units that are *always* dissimilar from each other. Using textual analysis, Hoberg and Phillips (2018) show that conglomerates often choose to operate in related industries – where there is a higher overlap between products. This conflicts with the linear (or logarithmic) relationship between the number of business units and operational complexity (e.g. Cohen and Lou 2012; Franco et al. 2016).

From a theoretical standpoint, firms that operate in several business units face a trade off between searching for synergies by choosing complementary (related) industries and choosing different (unrelated) industries in order to diversify away their risk (Hoberg and Phillips 2018). When using the number of business segments as a proxy for complexity, we assume that conglomerates represent a source of intra-industry heterogeneity regardless of the chosen strategy (synergies or diversification). Here, I adopt a different perspective and argue that conglomerates should be considered as heterogeneous (i.e. complex) only when they operate in unrelated business units. Since many financial statement users specialize in industries (e.g. analysts, asset managers), I argue that information processing costs are likely lower when conglomerates choose the synergy strategy over the diversification strategy. I argue that industry expertise may be more easily transferred to industries that are close to one's specialization than to industries that are "further away". Consequently, the relatedness (distance) between business segments should be considered in order to assess the overall complexity of a conglomerate. Thus, I suggest that the mere number of business segments is an incomplete proxy for operational complexity, as firms that share the same number of business segments could be considered either operationally complex or not, depending on the industries in which they operate.

To demonstrate this idea, I focus on financial analysts as a category of financial statements users. Financial analysts are sophisticated market participants that can mitigate the information asymmetry between investors and firms. Moreover, previous literature has highlighted that analysts specialize in industries in order to benefit from economies of scale when analyzing companies (Clement 1999; Ramnath 2002; Boni and Womack 2006; Bradley et al. 2017). For example, Boni and Womack (2006) show that analysts are good to rank firms inside industries, while Bradley et al. (2017) highlight the importance of prior work experience in the industry, before becoming an industry specialist as an analyst. More recently, the academic literature has focused on the analyst's skills outside their industry expertise. For example, Kadan et al. (2012) show that analysts not only possess within-industry expertise – the ability to rank firms inside industries – but also across-industry expertise. This expertise is useful for analysts to predict industry trends, even if the authors provide little information about where this expertise comes from. Also, Luo and Nagarajan

(2015) show that analysts can benefit from information complementarities between industries that are related through their supply chain.

On the other hand, the literature on complexity has investigated how analysts respond to both business and financial statement complexity. Multi-segment firms are usually seen as companies that have a complicated business model which requires more effort to understand their operations, leading to greater information processing costs for analysts (Dechow and You 2012). Some studies have modeled the association between analysts and complexity as a cost-benefit analysis (Lehavy et al. 2011). On the one hand, due to the higher information asymmetry between complex firms and investors, the demand for analyst services are higher, resulting in greater incentives for analysts to follow these firms. On the other hand, analysts must bear the increasing cost of processing complex information. Thus, they need to increase their effort (i.e., allocate more of their limited resources) to issue forecasts.

Based on the previous literature, I argue that firms operating in multiple (relatively) unrelated business units should be costlier to analyze than "regular" multiple-segment firms. Since analysts specialize in industries, I argue that the number of unrelated business units (hereafter *the number of complex business segments*) should make the earnings of the entire firm more difficult to forecast, leading to higher forecast dispersion and lower forecast accuracy. However, since analysts have both higher incentives and higher costs to cover these complex firms, I have no expectations regarding the sign of the association between the number of complex business segments and analyst coverage. Overall, I posit the following hypotheses:

H4.1: The number of complex business segments is not associated with analyst coverage.

H4.2: The number of complex business segments is negatively associated with analyst forecast accuracy.

H4.3: The number of complex business segments is positively associated with analyst forecast dispersion.

4.3. Research design

4.3.1. Construction of *CPX*

4.3.1.1. First step: Industry relatedness score

In this chapter, for each pair of SIC2 industries[38] I develop an industry relatedness score which represents the distance between two industries. Thus, industries that are farther (closer) from each other are classified as less (more) related. To evaluate the distance between industries, I assume that each industry occupies a unique location in an *n*-dimensional space. Dimensions are defined by fundamental ratios that characterize industries, consistent with prior literature on industry classifications (e.g., Bhojraj et al. 2003); each ratio corresponds to a separate dimension in the *n*-dimensional space. Operating ratios include capital intensity (*CAP_INT*), research and development expenses (*RD*), intangibles intensity (*XSGA*), cost of goods sold intensity (*COGS*), and asset turnover (*AT_TURN*). Profitability ratios include the two components of return on assets – cash flows from operations (*CFO*) and accruals (*TACC*) – and profit margin (*PM*). Valuation ratios include the book-to-market ratio (*BtoM*), enterprise value-to-sales (*EVS*) and the price-to-earnings ratio (*PE*). Finally, I add two ratios to reflect cross-industry differences in firms' financial structures: current ratio (*CR*) and leverage (*LEV*). Through this methodology, I assume that firms in *related* industries exhibit *related* fundamental ratios. Each year, to measure the distance between industries *i* and *j*, I calculate the Euclidean distance between both industries, *d*(*i*,*j*), using the following formula (time subscripts omitted):

$$d(i,j) = \sqrt{\sum_{v=1}^{n} (x_{iv} - x_{jv})^2} \tag{4.1}$$

[38] Previous literature (e.g. Bhojraj et al. 2003) highlights the superiority of GICS over the SIC classification in terms of intra-industry homogeneity. However, Compustat does not assign a GICS code for each business segment and only uses SIC or NAICS codes. Thus, I use the SIC classification to identify the number of complex business segments.

Where v indexes dimensions (ratios) in the n-space, n equals the number of dimensions (i.e. 13 ratios), and x_{iv} (x_{jv}) is the mean value of ratio v for single-segment firms in industry i (j).[39]

Ultimately, each SIC2 industry is mapped yearly in this spatial representation and has an industry relatedness score with all the other SIC2 industries.

4.3.1.2. Second step: Firm-level measure of complexity

From a theoretical standpoint, I assume that operationally complex firms are firms that have *complex* business units. Thus, I identify *complex* business segments as segments in an industry whose distance to the firm's primary industry is higher than the median distance between all industry pairs in that year. Moreover, I assume that the relationship between the number of complex segments and operational complexity is not linear (i.e. the marginal effect of one additional complex segment on complexity is a decreasing function). Consequently, I use the logarithmic transformation of *CPX* (*Log_CPX)* as my main variable of interest. Ultimately, the variable of interest *Log_CPX* is the log of one plus the number of complex business units. For example, let's take two firms operating primarily into the *sic28* "Chemicals and Allied Products" industry in 2014, Johnson & Johnson (gvkey=006266) and BASF (gvkey=017436). Both firms are multiple-segment firms and have several business units grouped in their primary industry (*sic28*). Moreover, BASF has a second business unit in the *sic13* industry *"Oil and Gas Extraction"* while Johnson & Johnson has a secondary business unit in the sic*38 "Instruments and Related Products"*. In Appendix D I show that for the year 2014 the distance between the primary industry (*sic28*) of these firms, and their secondary business segments belonging to the industry *sic13* and the industry *sic38* is respectively 3.70 and 1.24. The median distance across all the SIC2-pairs is 2.34 for the year 2014. Thus, every secondary segments that belong to *sic13* is considered as complex for firms belonging to the industry *sic28* (3.70>2.34). On the other hand, a secondary segment in the industry *sic38* is not considered complex since the industries *sic28* and *sic38* are closely

[39] To equalize the influence of all ratios on the aggregate distance measure, I standardize all ratios to a mean of zero and a standard deviation of one.

related (1.24<2.34). In the end, only BASF will have one complex business segment, and ultimately the variable *Log_CPX* will be equal to the natural log of 2 (i.e. 0.69).

4.3.2. Empirical models

To test my hypotheses, I use the same models as in chapter 3 regarding analysts forecast properties (dispersion, accuracy) and coverage[40]. However, for this chapter I add one additional control for financial statement complexity. I control for financial statement complexity since previous studies highlight its association with both analysts' forecasts and business complexity (Lehavy et al. 2011; Guay et al. 2016). I use the *Bog Index* as a proxy for financial statement complexity due to its theoretical and empirical superiority (Bonsall et al. 2017)[41].

4.3.3. Sample construction

Since I use both segment and firm level data, I construct two different samples. First, Table 4.1 – Panel A describes the sample that will be used in the empirical results section. Starting from the initial sample of 98,746 firm-year observations as explained in section 2.2.3, I delete firms with negative sales, total assets, or market capitalization less than 10 million, and firms with negative common equity. Finally, I delete firm-year observations with missing control variables, analyst, or Bog index data. Ultimately it leads to a sample of 21,378 observations, composed of 11,211 (10,167) multiple (single) segment firms. Panel B presents the sample attrition for the calculation of industry distance scores. Using the 54,228 firm-year observations from Panel A (before deleting missing control variables, analyst or Bog index data[42]), I then delete missing observations on the fundamental ratios, and keep only single-segment firms in industries (at SIC2 level) that have more than 5 observations per year. This leads to a final sample of 21,521 firm-year observations used to calculate the industry relatedness score. Finally, Panel C presents the sample attrition regarding segment-level data used to calculate the number of complex business units (*CPX*). First, I delete segments with missing industry information and missing sales and keep only segments with positive sales.

[40] See section 3.3.2.2 for more details about the models estimated.

[41] Chapter 3 did not include the Bog Index as a control because it would cause a significant drop in firm-year observations.

[42] Inversely, in Panel A I do not delete firms with missing observations on the fundamental ratios that are not used as control variables in the empirical results to maximize the number of observations.

Then, I merge segments at SIC2 level for each firm year resulting in 155,764 segment-year observations. It results in 102,771 segment-year observations after deleting segments where I was not able to calculate the distance with the primary business unit.[43]

TABLE 4.1: Sample selection

Panel A: Firm-year-level data

	Firm-year observations
Initial sample	98,746
Less firms with negative sales; assets under 10 millions; market capitalization under 10 millions; or negative common equity	(44,518)
Less missing control variables, analyst or Bog index data	(32,850)
Total number of observations in empirical tests	21,378
Including:	
Single-segment observations	10,167
Multiple segment observations	11,211

Panel B: Calculation of industry distance scores using single-segment firms

	Firm-year observations
Firms from Panel A (after exclusion of firms with negative sales; assets under 10 millions; market capitalization under 10 millions; or negative common equity)	54,228
Single-segment observations at SIC2 level	(22,777)
Less missing data for input variables of the calculation of the euclidean distance	(9,098)
Less industries with less than 5 observations per year	(832)
Single-segment observations used for calculation of industry distance scores	21,521

Panel C: Segment-level data

	Segment-year observations
US Segments in Compustat from 1999 to 2018	264,490
Less segments with missing industry information	(37,587)
Less segments with zero, negative or missing sales	(14,763)
Unique segments at SIC2-level	(56,376)
Less segments with missing industry distance	(52,993)
Segments usable for calculation of CPX	102,771

[43] This is mainly caused by segments that belongs to the financial industries (SIC60 to SIC69).

4.4. Results

4.4.1. Methodology results

Table 4.2 provides a summary of descriptive statistics regarding the industry relatedness score. First, Panel A presents the mean score and standard deviation for each industry. The mean industry relatedness score can be interpreted as how far an industry is different from the others on average. Thus, the industries *sic54 "Food Stores", sic13 "Oil and Gas extraction"* and *sic46 "Pipelines, except Natural Gas"* are the three industries exhibiting the highest average distance. Panel B provides more detailed results regarding the more (less) closely related pairs of industries. For example, in Panel B, the two most closely related industries are the industries *sic35 "Industrial Machinery and Equipment"* and *sic34 "Fabricated Metal Products, except Machinery"*. On the contrary, in Panel C I observe that the less related industries are the industries *sic54 "Food Stores"* and *sic13 "Oil and Gas Extraction"* since they have the highest distance.

TABLE 4.2: Industry relatedness score

Panel A: Mean distances

SIC2 code	Industry name	Mean	STD
54	Food Stores	3.65	1.06
13	Oil and Gas Extraction	3.62	1.06
46	Pipelines, except Natural Gas	3.59	1.12
70	Hotels and Other Lodging Places	3.31	1.10
10	Metal Mining	3.22	1.20
44	Water Transportation	3.08	1.01
55	Automotive Dealers & Service Stations	3.00	1.06
47	Transportations Services	2.87	1.03
58	Eating and Drinking Places	2.86	0.81
12	Bituminous Coal and Lignite Mining	2.82	0.82
14	Nonmetallic minerals, except fuels	2.75	0.88
79	Amusement and Recreation Services	2.74	0.93
45	Transportation by Air	2.72	0.87
31	Leather and Leather Products	2.69	1.03
78	Motion Pictures	2.68	0.81
28	Chemicals and Allied Products	2.67	0.87
22	Textile Mill Products	2.66	0.93
38	Instruments and Related Products	2.64	0.91
48	Communications	2.63	0.90
01	Agricultural Production - Crops	2.62	0.88
57	Furniture and Homefurnishings Stores	2.58	0.93
52	Building Materials & Garden Supplies	2.56	0.88
82	Educational Services	2.55	0.76
33	Primary Metal Industries	2.55	0.90
50	Wholesale Trade-Durable Goods	2.52	1.03
29	Petroleum Refining and Related Industries	2.52	0.85
56	Apparel & Accessory Stores	2.48	0.90
83	Social Services	2.42	0.83
53	General Merchandise Stores	2.42	0.90
49	Electric, Gas and Sanitary Services	2.42	0.87
59	Miscellaneous Retail	2.40	0.95
72	Personal Services	2.37	0.76
17	Construction Special Trade Contractors	2.37	0.82
51	Wholesale Trade-Nondurable Goods	2.36	0.94
23	Apparel and Other Textile Products	2.25	0.95
36	Electronic and Other Electric Equipment	2.22	0.87
16	Heavy construction, except building	2.20	0.75
39	Miscellaneous Manufacturing Industries	2.18	0.91
32	Stone, Clay, Glass, and Concrete Products	2.17	0.84
25	Furnitures and Fixtures	2.14	0.90
26	Paper and Allied Products	2.13	0.76
27	Printing, Publishing, and Allied Industries	2.11	0.80
73	Business Services	2.10	0.78
87	Engineering and Management Services	2.10	0.78
24	Lumber and Wood Products	2.07	0.79
35	Industrial Machinery and Equipment	2.00	0.88
30	Rubber and Miscellaneous Plastics Products	1.98	0.83
34	Fabricated Metal Products, except Machinery	1.97	0.88
37	Transportation Equipment	1.95	0.82
80	Health Services	1.91	0.76
20	Food and Kindred Products	1.85	0.78

Panel B: Top 10 most related industries

SIC2 code	Industry name	SIC2 code	Industry name	Mean
35	Industrial Machinery and Equipment	34	Fabricated Metal Products, except Machinery	0.75
34	Fabricated Metal Products, except Machinery	20	Food and Kindred Products	0.81
30	Rubber and Miscellaneous Plastics Products	20	Food and Kindred Products	0.84
80	Health Services	73	Business Services	0.89
37	Transportation Equipment	34	Fabricated Metal Products, except Machinery	0.91
34	Fabricated Metal Products, except Machinery	30	Rubber and Miscellaneous Plastics Products	0.91
87	Engineering and Management Services	73	Business Services	0.96
38	Instruments and Related Products	28	Chemicals and Allied Products	0.96
87	Engineering and Management Services	80	Health Services	0.96
37	Transportation Equipment	35	Industrial Machinery and Equipment	0.97

Panel C: Top 10 less related industries

SIC2 code	Industry name	SIC2 code	Industry name	Mean
54	Food Stores	13	Oil and Gas Extraction	5.31
55	Automotive Dealers & Service Stations	13	Oil and Gas Extraction	5.19
54	Food Stores	44	Water Transportation	5.10
50	Wholesale Trade-Durable Goods	13	Oil and Gas Extraction	4.97
47	Transportations Services	13	Oil and Gas Extraction	4.94
31	Leather and Leather Products	13	Oil and Gas Extraction	4.78
54	Food Stores	48	Communications	4.69
59	Miscellaneous Retail	13	Oil and Gas Extraction	4.63
51	Wholesale Trade-Nondurable Goods	13	Oil and Gas Extraction	4.63
54	Food Stores	38	Instruments and Related Products	4.63

This table presents results regarding the relatedness score estimated for each pair of SIC2 industry. In Panel A, I present the mean (standard deviation) distance for each industry with all the other industries. In Panel B and C, I compute the mean distance across years for each pair of industry. I require to have at least 10 observations (10 SIC2 pair-year) to include the mean in this table. In Panel B, I present the top 10 more related industries (smallest mean distance). In Panel C, I present the top 10 less related industries (highest mean distance).

4.4.2. Univariate statistics

Table 4.3 presents descriptive statistics for all firms (first three columns) and for multiple-segment firms (last three columns). The main variable of interest is *CPX* which represents the raw number of complex business units for each firm-year observation. On average, firms (multiple-segment firms) report 2.036 (2.976) business units (*NBS*), and operate in an average of 1.368 (1.701) SIC2 industries (*NBS_Sic2*); *NBS* and *NBS_Sic2* are set to one for single-segment firms but some firms with *NBS*>1 could still have *NBS_Sic2* equal to one if all its segments were concentrated in one SIC2 industry. The decline in the mean from *NBS* to *NBS_Sic2* suggests that many firms have multiple segments in the same SIC2 industry. Given that *NBS_Sic2* is at least 1, the mean value of *NBS_Sic2* implies that firms have on average 0.368 secondary business units in a SIC2 industry that is distinct from their primary industry. Among these secondary business units, the methodology developed here classifies 27% (0.101/0.368=27%) of them as complex business units (variable *CPX*). When looking exclusively at multi-segment firms (*NBS* > 1), I observe that these firms have on average 0.192 complex business units. Similarly to previous studies in the analyst's literature (e.g. Thomas 2002), I note that multi-segment firms receive more coverage from analysts, and their analysts' forecasts are less dispersed and more accurate than for single-segment firms. Finally, I observe that multi-segment firms are bigger since they have an average *Size* of (7.171 versus 6.698 for the full sample). They are also less volatile both in terms of return on assets (*σ(ROA)*) or stock returns (*σ(RET)*).

Table 4.4 presents the univariate correlation between variables used in the regression results[44]. Unsurprisingly, *CPX* has a strong positive correlation with *NBS* (0.31) and *NBS_Sic2* (0.48). Firms with complex segments (*CPX*) are larger (positive correlation of 0.13 with *Size*) and less volatile (negative correlation of -0.08 and -0.10 with *σ(ROA)* and *σ(RET)* respectively). Again, based on this correlation table, the correlations of complex business segments with forecast dispersion, forecast accuracy, and analyst coverage is also very weak.

[44] For simplification purpose, I restrict my analysis to Pearson correlations. Results using Spearman correlations yields similar inferences.

TABLE 4.3: Descriptive statistics

Sample	***Full***			***MS* (*NBS*>1)**		
Variable	**N**	**Mean**	**StdDev**	**N**	**Mean**	**StdDev**
NBS_Sic2	21378	1.368	0.684	11211	1.701	0.812
NBS	21378	2.036	1.276	11211	2.976	1.117
Log_NBS	21378	0.540	0.570	11211	1.029	0.340
CPX	21378	0.101	0.344	11211	0.192	0.455
Log_CPX	21378	0.066	0.217	11211	0.126	0.287
Size	21378	6.698	1.740	11211	7.171	1.692
BtoM	21378	0.565	0.484	11211	0.583	0.445
Volume	21378	4.284	1.564	11211	4.368	1.559
RD	21378	0.168	0.772	11211	0.029	0.180
Depreciation	21378	0.045	0.029	11211	0.042	0.024
SUE	21378	5.621	19.817	11211	5.872	19.994
Issue3y	21378	0.756	0.429	11211	0.830	0.376
σ(ROA)	21378	0.053	0.066	11211	0.040	0.050
σ(RET)	21378	-3.663	0.482	11211	-3.761	0.475
NegSUE	21378	0.416	0.493	11211	0.403	0.490
Loss	21378	0.252	0.434	11211	0.189	0.392
NegSI	21378	0.015	0.040	11211	0.016	0.039
Coverage	21378	1.620	0.958	11211	1.640	0.946
Accuracy	21378	-1.346	3.910	11211	-1.044	3.236
Dispersion	18277	0.417	0.930	9665	0.291	0.691
Bog	21378	4.441	0.085	11211	4.445	0.079

This table presents descriptive statistics for industry determinants and dependent variable from the regression results. The first three columns present results for the full sample. The last three columns present results for multiple-segment firms only (*MS*). Multiple-segment firms are firms with *NBS*>1. All variables are winsorized at 1% and 99%, except for Accuracy and Dispersion winsorized at 5% and 95%.

Finally, the correlation between *CPX* and financial statement complexity proxied by the Bog index is insignificant and almost zero. This is surprising since business complexity is considered as one of the two primary determinants of financial statement complexity (Guay et al. 2016). However, previous literature shows that managers have incentives to obfuscate information and issue complex financial statements regardless of the complexity of the underlying economics (e.g. Bushee et al. 2018). This could result in a noisy relationship between business complexity and financial statement complexity.

TABLE 4.4: Univariate correlations

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)
(1)	NBS		0.62	0.31	0.37	0.02	0.13	-0.15	-0.09	0.04	0.18	-0.20	-0.24	-0.03	-0.14	0.00	0.09	0.09	-0.13	0.06
(2)	NBS_Sic2	0.63		0.48	0.23	0.00	0.07	-0.10	-0.07	0.03	0.13	-0.13	-0.17	-0.03	-0.10	-0.01	0.04	0.06	-0.08	0.01
(3)	CPX	0.30	0.46		0.13	0.00	0.04	-0.05	0.01	0.03	0.07	-0.08	-0.10	-0.02	-0.05	-0.02	0.04	0.03	-0.03	0.01
(4)	Size	0.35	0.22	0.13		-0.01	0.69	-0.21	-0.05	0.20	0.35	-0.37	-0.48	-0.03	-0.26	-0.05	0.64	0.21	-0.21	0.02
(5)	BtoM	0.10	0.05	0.03	0.02		-0.18	-0.09	0.07	0.09	0.07	-0.01	0.30	0.18	0.18	0.13	-0.27	-0.22	0.13	-0.07
(6)	Volume	0.10	0.06	0.04	0.68	-0.24		0.04	-0.01	0.25	0.13	0.01	-0.13	-0.02	-0.03	0.03	0.73	0.09	-0.01	0.16
(7)	RD	-0.07	-0.05	-0.10	-0.24	-0.26	0.04		-0.13	-0.01	-0.10	0.27	0.22	0.10	0.33	0.04	-0.03	-0.12	0.22	0.27
(8)	Depreciation	-0.06	-0.04	0.02	-0.01	0.05	-0.01	-0.21		0.07	0.07	0.05	0.13	0.03	0.09	0.02	0.00	-0.06	0.09	-0.18
(9)	SUE	0.07	0.04	0.03	0.32	0.09	0.45	0.02	0.06		0.07	0.22	0.08	0.00	0.16	0.16	0.08	-0.17	0.21	0.05
(10)	Issue3y	0.20	0.13	0.07	0.36	0.10	0.14	-0.16	0.09	0.12		-0.17	-0.13	0.02	-0.06	0.01	0.13	0.02	-0.03	0.00
(11)	σ(ROA)	-0.22	-0.15	-0.09	-0.41	-0.03	-0.02	0.27	0.06	0.40	-0.19		0.39	0.09	0.41	0.30	-0.17	-0.27	0.32	0.19
(12)	σ(RET)	-0.25	-0.17	-0.10	-0.49	0.18	-0.14	0.14	0.08	0.14	-0.14	0.47		0.15	0.45	0.22	-0.32	-0.32	0.32	0.05
(13)	NegSUE	-0.03	-0.03	-0.02	-0.03	0.17	-0.02	0.04	0.03	0.06	0.02	0.14	0.13		0.31	0.27	-0.07	-0.09	0.08	0.05
(14)	Loss	-0.16	-0.11	-0.05	-0.27	0.08	-0.04	0.24	0.05	0.32	-0.06	0.46	0.44	0.31		0.38	-0.20	-0.32	0.37	0.22
(15)	NegSI	0.10	0.04	-0.01	0.12	0.07	0.12	0.11	0.03	0.23	0.10	0.16	0.05	0.28	0.27		-0.07	-0.18	0.09	0.08
(16)	Coverage	0.07	0.04	0.04	0.65	-0.28	0.74	-0.03	0.01	0.19	0.13	-0.20	-0.33	-0.07	-0.20	0.03		0.27	-0.20	0.05
(17)	Accuracy	0.10	0.08	0.04	0.29	-0.26	0.22	-0.04	-0.05	-0.22	0.03	-0.34	-0.37	-0.14	-0.38	-0.10	0.41		-0.67	-0.06
(18)	Dispersion	-0.12	-0.08	0.00	-0.19	0.24	-0.06	0.04	0.08	0.32	0.00	0.38	0.38	0.17	0.43	0.10	-0.24	-0.60		0.14
(19)	Bog	0.08	0.03	0.00	0.02	-0.09	0.15	0.38	-0.21	0.12	0.00	0.14	0.04	0.04	0.22	0.12	0.05	-0.09	0.15	

This table presents the univariate correlations for the main variables. The top (bottom) of the table represents Pearson (Spearman) correlations.

4.4.3. Regression analysis

4.4.3.1. Analyst coverage

Results are presented in Table 4.5. Panel A presents the results for the full sample, while Panel B outlines the results using multiple-segments firms exclusively (*NBS*>1). Using the full sample, in model (1) *Log_CPX* is negatively associated with coverage (coefficient of -0.1205; p < 0.01). The results are similar when *Log_CPX* is replaced with *Log_NBS* (model (2)). However, when both variables are included (model (3)), only the coefficient on *Log_NBS* remains significant. Panel B confirms these results (except for model (1) where the coefficient on *Log_CPX* is no longer significant). Overall, the evidence suggests that analysts may be reluctant to follow firms that have several business units, regardless of whether these segments are related to the firm's primary business. All control variables are significant and consistent with prior literature. Finally, I confirm that financial statement complexity (as proxied by the *Bog index*) is negatively associated with coverage.

TABLE 4.5: Analyst coverage

Panel A: Full sample

Sample		***Full***			***Full***			***Full***	
Model		***(1)***			***(2)***			***(3)***	
	Coeff.	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
Log_CPX	**-0.1205**	*******	**-3.6**				**0.0068**		**0.2**
Log_NBS				**-0.1659**	*******	**-9.8**	**-0.1666**	*******	**-9.2**
Bog	-0.3196	**	-2.3	-0.2108	*	-1.6	-0.2108	*	-1.6
Size	0.1136	***	8.9	0.1353	***	10.4	0.1353	***	10.4
BtoM	-0.2704	***	-7.0	-0.2712	***	-7.2	-0.2713	***	-7.2
Volume	0.3419	***	30.9	0.3301	***	29.8	0.3301	***	29.8
R&D	0.0363	***	2.9	0.0247	**	2.0	0.0247	**	2.0
Depreciation	1.5342	***	5.1	1.2653	***	4.3	1.2634	***	4.3
Issue_3y	-0.0664	***	-2.9	-0.0559	**	-2.5	-0.0559	**	-2.5
σ(ROA)	-1.1648	***	-8.5	-1.1889	***	-8.9	-1.1888	***	-8.9
σ(RET)	-0.1843	***	-6.7	-0.1861	***	-6.7	-0.1861	***	-6.7
Intercept	0.3361		0.5	-0.1592		-0.3	-0.1591		-0.3
Adjusted R2	0.61			0.62			0.62		
N	21,378			21,378			21,378		

Panel B: Multiple-segments firms only

Sample		***MS***			***MS***			***MS***	
Model		**(1)**			**(2)**			**(3)**	
	Coeff.	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
Log_CPX	**-0.0198**		**-0.6**				**0.0006**		**0.2**
Log_NBS				**-0.0894**	*******	**-2.7**	**-0.0894**	*******	**-2.7**
Bog	-0.3982	***	-2.6	-0.3837	**	-2.5	-0.3837	**	-2.5
Size	0.1110	***	6.3	0.1190	***	6.6	0.1190	***	6.6
BtoM	-0.2657	***	-6.3	-0.2681	***	-6.3	-0.2681	***	-6.3
Volume	0.3398	***	21.1	0.3368	***	20.8	0.3368	***	20.7
R&D	0.0071		0.2	0.0022		0.0	0.0022		0.0
Depreciation	1.1082	**	2.4	1.1047	**	2.3	1.1047	**	2.3
Issue_3y	-0.0536	**	-2.1	-0.0536	**	-2.1	-0.0536	**	-2.1
σ(ROA)	-1.4551	***	-7.0	-1.4332	***	-7.0	-1.4332	***	-7.0
σ(RET)	-0.1797	***	-4.8	-0.1784	***	-4.7	-0.1784	***	-4.7
Intercept	0.6675		1.0	0.6543		1.0	0.6543		1.0
Adjusted R2	0.63			0.63			0.63		
N	11,211			11,211			11,211		

This table reports coefficients estimates (Coeff.), statistical significance (Sig.) and t-statistics (t.) from a regression of analyst coverage (Coverage) on Log_CPX and control variables. Panel A presents results for the full sample. Panel B presents the result for multiple-segments firms only (NBS>1). In the Sig. column, a *** (**; *) indicates that the coefficient is different from zero at the 1% (5%; 10%) level (two-tailed). Standard errors are double-clustered by firm and time.

4.4.3.2. Analyst forecast accuracy

Table 4.6 presents the results regarding forecast accuracy. In Panel A, results using the full sample indicate that both proxies for operational complexity (*Log_NBS* and *Log_CPX*) have negative effects on forecast accuracy, as they take negative coefficients in the first two specifications. In model (3) both coefficients become insignificant[45]. However, when the sample is restricted to multi-segment firms (Panel B), only *Log_CPX* remains negative and significant at the 10% level in model (3). The coefficient on *Log_NBS* is no longer significant in model (2) as well. Thus, while the results from Panel A suggest that the *existence* of multiple segments may trigger a decline in forecast accuracy, the results with or without single-segment firms indicate that the *complexity* of the additional business segments is

[45] As highlighted before (see Table 4.4) *Log_NBS and Log_CPX* exhibit a high positive correlation. This could create collinearity problem and increase the standard errors of the coefficients.

associated with forecast accuracy. All other control variables, including a proxy for financial statement complexity, are significant in accordance with previous literature.

TABLE 4.6: Analyst forecast accuracy

Panel A: Full sample

Sample		***Full***			***Full***			***Full***	
	Coeff.	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
Log_CPX	**-0.2514**	**	**-2.2**				**-0.1797**		**-1.3**
Log_NBS				**-0.1153**	*	**-1.6**	**-0.0918**		**-1.1**
Bog	0.0161		0.0	0.0739		0.1	0.0675		0.1
Size	0.1741	***	5.2	0.1815	***	5.4	0.1816	***	5.4
σ(ROA)	-4.3128	***	-3.1	-4.3755	***	-3.2	-4.3816	***	-3.2
σ(RET)	-1.2459	***	-8.1	-1.2501	***	-8.4	-1.2520	***	-8.4
SUE	-0.0255	***	-7.2	-0.0256	***	-7.2	-0.0255	***	-7.2
NegSUE	0.0823		1.1	0.0814		1.1	0.0812		1.1
Loss	-1.5302	***	-7.7	-1.5386	***	-7.8	-1.5371	***	-7.8
NegSI	-4.1958		-1.3	-4.0675		-1.3	-4.0895		-1.3
Days	-0.3847	***	-5.6	-0.3832	***	-5.5	-0.3835	***	-5.5
Intercept	-5.3236	**	-2.4	-5.5987	**	-2.5	-5.5779	**	-2.5
Adjusted R2	0.18			0.18			0.18		
N	21,378			21,378			21,378		

Panel B: Multiple-segments firms only

Sample		***MS***			***MS***			***MS***	
	Coeff.	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
Log_CPX	**-0.1979**	*	**-1.6**				**-0.1888**	*	**-1.6**
Log_NBS				**-0.0736**		**-0.6**	**-0.0398**		**-0.3**
Bog	-1.6250	***	-2.6	-1.6041	***	-2.6	-1.6200	***	-2.6
Size	0.1958	***	5.9	0.1987	***	5.8	0.1983	***	5.8
σ(ROA)	-6.1643	***	-3.6	-6.1270	***	-3.6	-6.1632	***	-3.6
σ(RET)	-0.8258	***	-6.2	-0.8229	***	-6.2	-0.8263	***	-6.3
SUE	-0.0201	***	-5.0	-0.0202	***	-5.0	-0.0201	***	-5.0
NegSUE	0.0826		0.8	0.0825		0.8	0.0822		0.8
Loss	-1.6900	***	-8.8	-1.6924	***	-8.8	-1.6902	***	-8.8
NegSI	-1.6546		-0.6	-1.6131		-0.6	-1.6468		-0.6
Days	-0.3475	***	-5.0	-0.3466	***	-5.0	-0.3474	***	-5.0
Intercept	3.2867		1.4	3.2314		1.3	3.2838		1.4
Adjusted R2	0.19			0.19			0.19		
N	11,211			11,211			11,211		

This table reports coefficients estimates (Coeff.), statistical significance (Sig.) and t-statistics (t.) from a regression of analyst forecast accuracy (Accuracy) on Log_CPX and control variables. Panel A presents results for the full sample. Panel B presents the result for multiple-segments firms only (NBS>1). In the Sig. column, a *** (**; *) indicates that the coefficient is different from zero at the 1% (5%; 10%) level (two-tailed). Standard errors are double-clustered by firm and time.

4.4.3.3. Analyst forecast dispersion

Table 4.7 presents the results regarding the analyst forecast dispersion. Using the full sample (Panel A) operational complexity is not significantly associated with the forecast dispersion in model (1). In contrast, in model (2), which replaces *Log_CPX* with the log number of business units (*Log_NBS*), this latter proxy for operational complexity is surprisingly negatively associated with forecast dispersion. Finally, when both *Log_NBS* and *Log_CPX* are included (model (3)), only *Log_CPX* presents a significant positive association at 5 % level with the forecasts' dispersion (coefficient on *Log_CPX*: 0.0780) while *Log_NBS* remains negatively associated with forecast dispersion (at 1% level). Financial statement complexity (*BOG*) is positively associated with dispersion, confirming results provided by previous literature (Lehavy et al. 2011). All other control variables are significant at 5% or 1% level, except for *NegSUE*. In panel B, *Log_CPX* is always positively associated with forecast dispersion (at 5% level). In contrast, the coefficient on *Log_NBS* is insignificant in model (2) and (3). These results suggest that only the *complexity* of business units is associated with analyst forecast dispersion. Overall, these results are consistent with H4.3.

TABLE 4.7: Analyst forecast dispersion

Panel A: Full sample

Sample		***Full***			***Full***			***Full***	
	Coeff.	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
Log_CPX	**0.0363**		**1.2**				**0.0780**	**	**2.3**
Log_NBS				**-0.0437**	**	**-2.2**	**-0.0540**	***	**-2.6**
Bog	0.5629	***	3.7	0.5913	***	3.8	0.5939	***	3.8
Size	-0.0399	***	-3.7	-0.0350	***	-3.4	-0.0350	***	-3.4
σ(ROA)	1.6576	***	4.7	1.6121	***	4.7	1.6152	***	4.7
σ(RET)	0.2811	***	6.6	0.2768	***	6.6	0.2776	***	6.6
SUE	0.0076	***	6.5	0.0076	***	6.6	0.0076	***	6.6
NegSUE	-0.0153		-0.6	-0.0160		-0.7	-0.0161		-0.7
Loss	0.5448	***	9.2	0.5409	***	9.3	0.5402	***	9.3
NegSI	-2.2680	***	-7.7	-2.2112	***	-7.6	-2.2017	***	-7.7
Days	0.0204	**	2.5	0.0206	**	2.5	0.0207	**	2.6
Intercept	-1.0357	*	-1.7	-1.1826	*	-1.9	-1.1913	*	-1.9
Adjusted R2	0.22			0.22			0.22		
N	18,277			18,277			18,277		

Panel B: Multiple-segments firms only

Sample		*MS*			*MS*			*MS*	
	Coeff.	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**	**Coeff.**	**Sig.**	**t.**
Log_CPX	**0.0676**	**	**2.1**				**0.0720**	**	**2.2**
Log_NBS				**-0.0065**		**-0.3**	**-0.0195**		**-0.8**
Bog	0.5550	***	3.7	0.5518	***	3.7	0.5573	***	3.7
Size	-0.0380	***	-4.2	-0.0368	***	-4.2	-0.0367	***	-4.2
σ(ROA)	1.5118	***	4.0	1.4963	***	4.0	1.5116	***	4.0
σ(RET)	0.1855	***	4.4	0.1841	***	4.4	0.1852	***	4.4
SUE	0.0068	***	7.2	0.0068	***	7.2	0.0068	***	7.2
NegSUE	-0.0061		-0.3	-0.0063		-0.3	-0.0063		-0.3
Loss	0.4488	***	9.2	0.4498	***	9.2	0.4488	***	9.2
NegSI	-1.7925	***	-4.2	-1.8000	***	-4.2	-1.7887	***	-4.2
Days	0.0031		0.3	0.0028		0.3	0.0031		0.3
Intercept	-1.3482	**	-2.2	-1.3310	**	-2.2	-1.3496	**	-2.2
Adjusted R2	0.20			0.20			0.20		
N	9,655			9,665			9,655		

This table reports coefficients estimates (Coeff.), statistical significance (Sig.) and t-statistics (t.) from a regression of analyst forecast dispersion (*Dispersion*) on *Log_CPX* and control variables. Panel A presents results for the full sample. Panel B presents the result for multiple-segments firms only (*NBS*>1). In the Sig. column, a *** (**; *) indicates that the coefficient is different from zero at the 1% (5%; 10%) level (two-tailed). Standard errors are double-clustered by firm and time.

4.5. Summary

In this chapter I develop a new methodology to measure operational complexity. While proxies for operational complexity are typically based on the number of business units of the firms, I add a new dimension by considering the added complexity triggered by the inclusion of a business segment that operates in an industry that is unrelated to the firm's primary industry. I develop a new methodology to measure the inter-industry distance – the industry relatedness – using fundamental ratios. This methodology is used to create a new proxy for operational complexity and respond to the call made by Hoitash et Hoitash(2018) for more studies related to this topic.

Then, I test this measure in the context of financial analysts. I show that my measure is negatively (positively) associated with analyst forecast accuracy (dispersion), consistent with the argument that my measure captures a novel dimension of operational complexity. However, I find no association between my measure and analyst coverage after controlling for the number of business units, which indicates that analysts do not distinguish between relatively complex and relatively simpler multiple-segment firms in the decision to initiate or discontinue coverage. This result is surprising given that earlier results suggested that complexity is a significant driver of forecast dispersion and accuracy. In other words, although analysts may choose not to cover multiple-segment firms because these generally have earnings that are more difficult to forecast, once coverage is initiated, analysts are less successful in predicting the earnings of the subset of multiple-segment firms that have complex business segments.

Despite controlling for some confounding effects and testing the robustness of my results through several models, my study still suffers from some limitations. First, I subjectively choose the fundamental ratios used to calculate the distance between industries. One could argue that they do not represent all the relevant dimensions on which industry relatedness should be measured. The correct set of fundamentals is ultimately an empirical issue; to the extent that industry distance is measured with error, the results on the association between complexity and analyst forecast properties may be understated. Also, I acknowledge that only using single-segment firms to compute the distance between industries may not always be appropriate. Some industries may be characterized by *typical* multiple-segment firms rather than single-segment firms. Therefore, future studies could investigate whether using single-segment firms systematically as a representative sample of industries is correct. In addition, my measure of operational complexity contributes to the whole debate of how complexity impact financial statement users, and how operational and financial statement complexity interact with each other. Previous literature emphasizes how operational complexity and financial statement complexity should be associated. However, I do not consider this association. My new measure opens new avenues for future research which could investigate more closely the relationship between these two forms of complexity.

Conclusion

The general objective of this thesis is to document several sources of intra-industry heterogeneity and to investigate its implications. First, I find three sources that rely on the weaknesses and strengths of the use of industry classifications as a peer selection method.

In chapter 2, I build on the assumptions behind the creation and the use of the SIC and GICS and exploit the different dimensions of heterogeneity they represent to identify *industry classification misfits*. In the third chapter, I exploit the construction of industry classification that categorizes firms in closed boxes, and which assumes the transitivity of industry membership. I rely on firms' motivations to become differentiated and provide a continuous measure of intra-industry homogeneity at the firm-level which permits the identification of *differentiated firms*. Finally, in chapter 4 I exploit multi-segment firms as a natural source of intra-industry heterogeneity. Using the industry relatedness, I show that contrary to previous studies, multi-segment firms are not always a source of heterogeneity. Only *complex conglomerates* represent heterogeneous firms.

TABLE C.1: Comparisons between heterogeneity measures

Industry classification misfits										
	DIFF									
	1	2	3	4	5	6	7	8	9	10
MISFIT = 0 and belonging to the MOST frequent SIC2	63%	62%	61%	59%	58%	57%	55%	51%	50%	46%
MISFIT = 0 and **not** belonging to the MOST frequent SIC2	28%	27%	29%	29%	31%	31%	32%	33%	34%	35%
Misfits firms	9%	11%	10%	12%	12%	13%	13%	16%	16%	19%

Complex conglomerates										
	DIFF									
	1	2	3	4	5	6	7	8	9	10
Multiple-segments firms (NBS>1)	58%	58%	58%	53%	51%	53%	50%	52%	49%	45%
Mean NBS	2.143	2.136	2.106	2.072	2.013	2.017	1.910	1.973	1.891	1.774
Mean CPX	0.132	0.100	0.101	0.088	0.077	0.078	0.068	0.086	0.079	0.060

This table present descriptive comparisons between heterogeneity measures. For each differentiation deciles (DIFF), I present descriptive statistics regarding their link with both *industry classification misfits* and *complex conglomerates*.

One concern that arises in the light of this thesis is that these three sources of heterogeneity may catch up the same heterogeneous firms, even though they are conceptually and empirically different. Thus, in table C.1, I present some descriptive statistics to show how they differ[46]. First, for firms belonging to each differentiation deciles (DIFF), I expose how frequently they are also classified as industry core firms or misfits. I separate industry core firms in two categories: firms belonging to the most frequent GICS6-SIC2 or not. This subdivision allows identifying firms that are closer to the industry core. Results in table C.1 confirm this prediction since industry core firms (MISFIT = 0) possessing the most frequent GICS6-SIC2 combination are more recurrent in low differentiation deciles (63% of firms in DIFF=1) than in high differentiated deciles (46% of firms in DIFF=9). On the contrary, I notice that misfit firms (MISFIT=1) tend to be more present in higher differentiation deciles. This evidence suggests that there is an overlap between these two measures of intra-industry heterogeneity. Then, I present comparisons between differentiated firms and complex conglomerates. Through Table C.1, I observe that differentiated firms are more likely to be single-segment firms. In addition, differentiated firms possess less complex secondary business units. These results might be surprising even if it confirms the (untabulated) negative correlation between CPX and DIFF. One plausible explanation for this negative correlation is the possible mechanical link between differentiated firms and single-segment firms. In chapter 3, more volatile firms tend to be classified as differentiated since they have financial ratios that should be less correlated with their industry peers. On the other hand, complex conglomerates are probably firms that engage in a diversification strategy to decrease their risk through the reduction of the volatility of their operations. Thus, they are less likely to be classified as differentiated firms according to the methodology developed in chapter 3. In the end, it results in differentiated firms and complex conglomerates that seems to be a distinct type of heterogeneous firms. Overall, through Table C.1, I observe that the overlap between the three types of heterogeneous firms is marginal, which reinforces the interest of these three empirical measures of heterogeneity.

Then, I test the implications of intra-industry heterogeneity for several stakeholders. I show that industry classification misfits can impact the ability of researchers to predict misstatements through abnormal accruals. Also, I show that differentiated firms suffer from an informational cost on the stock market materialized through both investors and financial

[46] Additionally, in untabulated results I note that the correlation between MISFIT and DIFF (CPX) stands at 0.08 (0.09), while the correlation between DIFF and CPX is negative (-0.05). These untabulated results support the idea that they represent different types of heterogeneous firms.

analysts. Finally, I find that analysts have more difficulties to forecast complex conglomerates earnings compared to non-complex conglomerates.

This thesis makes four main contributions. First, it contributes to the literature on the use of industry classifications. The use of industry classifications as a peer selection method has been challenged recently by several studies (Lee et al. 2015; Hoberg and Phillips 2016; Hoberg and Phillips 2010; Ecker et al. 2013; Ding et al. 2019). However, industry classifications remain widely used both in research and practice because of the lack of replicability (or data availability) of alternative classifications, or because these alternatives rely on subjective criteria that diminish their value outside the context of their studies. In my thesis, I provide new insights regarding the use of industry classifications by shedding the light on three sources of intra-industry heterogeneity.

Second, I document the implications of intra-industry heterogeneity for practitioners. Many users of financial information (i.e. financial analysts, investors) use benchmarks to analyze firms and specialize in industries to benefit from economies of scale when analyzing firms. In this thesis, I show that being heterogeneous as a firm can come at a cost on financial markets. Firms may suffer from an informational cost on the stock market, while it may be costlier for financial statement users to analyze such firms. Thus, this thesis contributes to the literature on both asset pricing and financial analysts.

Third, I contribute to the literature on peer selection methods. Even if I do not formally propose a new methodology to classify firms into homogeneous groups, I document new methods to measure intra-industry homogeneity and inter-industry relatedness. These methods are easy to implement and to replicate since they use publicly available data. Thus, they could be used in the future to better select peers and to form more homogeneous groups of firms.

Finally, I contribute to the literature on accrual models. Previous studies show that the intra-industry homogeneity is unlikely to hold and may have implications for the estimation of accrual models (Peterson et al. 2015; Owens et al. 2017). I confirm these results by documenting a new source of intra-industry heterogeneity. Thus, I respond to the calls for a better understanding regarding how fundamentals can drive earnings quality measurement (Dechow et al. 2010).

Due to its exploratory nature, my thesis suffers from several limitations. First, I identify three sources of heterogeneity without formally comparing them. As shown in Table C.1, I acknowledge that a firm can potentially be simultaneously an *industry classification misfit,* a *differentiated firm,* and a *complex conglomerate.* However, this is not systematically the case since the methodologies leading to the identification of these three types of heterogeneous firms rely on different assumptions. Future research could provide more empirical evidence regarding this concern and highlight whether one or the other form of heterogeneous firms dominates.

Moreover, I voluntarily decide to provide empirical results specific to each chapter. More precisely, I do not include any empirical results using accruals models in chapters 3 and 4. In these chapters, I mainly use fundamental ratios to identify *differentiated firms* or *complex conglomerates*. These ratios include proxies for performance, volatility and liquidity which are highly correlated with accruals. Thus, I believe that making inferences regarding accruals of *differentiated firms* or *complex conglomerates* could be risky. Ultimately, it would lead to controversial contributions to the literature that I did not want to be included in this thesis. On the other hand, I did not provide any results regarding the consequences of *industry classification misfits* on financial analysts or information asymmetry. This choice is debatable since they represent a form of heterogeneous firms that could be affected in a similar way as *differentiated firms*. However, for parsimony reasons I decided to focus on accruals models and industry news incorporation in chapter 2. Moreover, *industry classification misfits* are identified exclusively through two industry classifications. Hypothesizing that financial analysts are affected by *industry classification misfits* implies that they actually use the GICS and SIC classifications to build their portfolio. Even if literature supports this idea (Boni and Womack 2006), I found this assumption too strong to support the potential results. Also, during my thesis I did not collect any data on financial analysts at the analyst-level. Future research using this type of data could investigate more closely the relationship between analyst portfolio and *industry classification misfits*.

Also, I always try to identify sources of heterogeneity departing from either the SIC or the GICS industry classifications. However, this choice remains debatable since alternative classifications seem to offer a better intra-industry homogeneity (Krishnan and Press 2003; Hoberg and Phillips 2016). I exclusively depart from these two classifications since they are heavily used in both practice and academia. Also, I do not think that my contributions are limited to the scope of the SIC or GICS. On the contrary, I think that applying the methodologies developed in this thesis to alternative classifications could yield interesting results and should be of interest of future research.

Finally, in chapters 3 and 4, I arbitrarily chose the fundamental ratios that could be categorized as industry characteristics. Although they are based on previous literature, one could argue that they do not fully represent every dimension of intra-industry heterogeneity. However, the main strength of the methodologies of these chapters is that they can be easily adapted. Thus, future research could add (or delete) dimensions to better identify intra-industry heterogeneity.

This thesis opens avenues for future research. First, the three methodologies provided to identify sources of intra-industry are easy to implement and could be further used to better select peers in many contexts. For example, in chapter 4 I provide a measure of distance for each pair of industries. This proxy for industry relatedness can serve to merge industries. This could be particularly useful when industries have a small number of observations that exclude them from studies (e.g. accruals models, international studies, etc.). In addition, in chapter 3 I develop a methodology that creates a firm-year measure of differentiation, based on the Euclidean distance as well. This methodology could be easily adapted to compute the distance between each pair of firms. Thus, it would provide a way to form peer groups of any sizes.

Also, in this thesis I focus on industry classifications. I believe that more research is needed to better understand who are using them, and more importantly how they are used. For example, the GICS is the classification that is the most frequently updated. How firms and financial statement users are affected by these changes could constitute an interesting research question. More precisely, in 2018 the MSCI processed a major revision of the communication services sector and the information technology sector leading to the reclassification of many firms. This setting could be fruitful to study the effect of industry classifications on firms and users.

Moreover, few studies focus on the consequences of heterogeneous for financial statements users. In this thesis I provide preliminary evidence regarding the economic consequences of being heterogeneous. However, the scope of this thesis is limited to the negative consequences for firms or financial statement users. For example, I highlight that differentiated firms suffer from higher information processing costs. Firms should react and implement actions to decrease these costs. More research is needed to understand how differentiated firms do so. Moreover, in this thesis I focus on the negative consequences of heterogeneity. Beyond the obvious strategic advantage it provides, future research could investigate what the other benefits associated with being a heterogeneous firm are.

I exclusively use industry classifications and fundamental ratios to identify heterogeneous firms. More dimensions to heterogeneity could be added in future research. For example, in chapter 4 I focus on operational complexity through the interpretation of *business* segments. The *geographical* segments could also be interpreted as a source of heterogeneity. Geographical diversification strategies seem more and more present in the global economy context we experience currently. Thus, firms' heterogeneous geographical locations could be an interesting topic to investigate.

Finally, I show that financial analysts have more difficulties to analyze heterogeneous firms. Still, some analysts are willing to follow them. Future research could investigate what are their incentives to better understand why analysts decide to cover these firms. Data collection at the analyst level could be particularly useful in that context.

Bibliography

(FASB), F. A. S. B. 1980. Qualitative Characteristics of Accounting Information. *Statement of Financial Accounting Concepts No. 2. Norwalk, CT: FASB, 1980.*

Ai, C., and E. C. Norton. 2003. Interaction terms in logit and probit models. *Economics Letters* 80 (1): 123–129.

Albuquerque, A. 2009. Peer firms in relative performance evaluation. *Journal of Accounting and Economics* 48 (1): 69–89.

Ali, U., and D. Hirshleifer. 2020. Shared analyst coverage: Unifying momentum spillover effects. *Journal of Financial Economics* 136 (3): 649–675.

Amihud, Y. 2002. Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets* 5 (1): 31–56.

Basu, S. 1997. The conservatism principle and the asymmetric timeliness of earnings. *Journal of Accounting and Economics* 24 (1): 3–37.

Bhojraj, S., C. M. C. Lee, and O. D. 2003. What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research* 41 (5): 745–774.

Boni, L., and K. L. Womack. 2006. Analysts, Industries, and Price Momentum. *Journal of Financial and Quantitative Analysis* 41 (1): 85–109.

Bonsall, S. B., A. J. Leone, B. P. Miller, and K. Rennekamp. 2017. A plain English measure of financial reporting readability. *Journal of Accounting and Economics* 63 (2–3): 329–357.

Bradley, D., S. Gokkaya, and X. Liu. 2017. Before an Analyst Becomes an Analyst: Does Industry Experience Matter? *Journal of Finance* 72 (2): 751–792.

Brown, L. D., A. C. Call, M. B. Clement, and N. Y. Sharp. 2015. Inside the "Black Box" of sell-side financial analysts. *Journal of Accounting Research* 53 (1): 1–47.

Bushee, B. J., I. D. Gow, and D. J. Taylor. 2018. Linguistic Complexity in Firm Disclosures: Obfuscation or Information? *Journal of Accounting Research* 56 (1): 85–121.

Cairney, T. D., and G. R. Young. 2006. Homogenous industries and auditor specialization: An indication of production economies. *Auditing: A Journal of Practice & Theory* 25 (1): 49–67.

Chen, H., L. Cohen, and D. Lou. 2016. Industry window dressing. *Review of Financial Studies* 29 (12): 3354–3393.

Chen, W., P. Hribar, and S. Melessa. 2018. Incorrect Inferences When Using Residuals as Dependent Variables. *Journal of Accounting Research* 56 (3): 751–796.

Choi, J. H., J. B. Kim, and Y. Zang. 2010. Do abnormally high audit fees impair audit quality? *Auditing: A Journal of Practice & Theory* 29 (2): 115–140.

Chun, H., J.-W. Kim, R. Morck, and B. Yeung. 2008. Creative destruction and firm-specific performance heterogeneity. *Journal of Financial Economics* 89 (1): 109–135.

Clarke, R. N. 1989. SICs as Delineators of Economic Markets. *The Journal of Business* 62 (1): 17–31.

Clement, M. B. 1999. Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter? *Journal of Accounting and Economics* 27 (3): 285–303.

Cohen, L., and A. Frazzini. 2008. Economic links and predictable returns. *Journal of Finance* 63 (4): 1977–2011.

Cohen, L., and D. Lou. 2012. Complicated firms. *Journal of Financial Economics* 104 (2): 383–400.

Dechow, P., and I. D. Dichev. 2002. The Quality of Accruals and Earings: The Role of Accruals Estimation Errors. *The Accounting Review* 77: 35–59.

Dechow, P., W. Ge, and C. Schrand. 2010. Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of Accounting and Economics* 50 (2–3): 344–401.

Dechow, P. M., W. Ge, C. R. Larson, and R. G. Sloan. 2011. Predicting Material Accounting Misstatements. *Contemporary Accounting Research* 28: 17–82.

Dechow, P. M., R. G. Sloan, and A. P. Sweeney. 1995. Detecting Earnings Management. *The Accounting Review* 70 (2): 193–225.

Dechow, P. M., and H. You. 2012. Analysts' motives for rounding EPS forecasts. *Accounting Review* 87 (6): 1939–1966.

Demski, J. J. S. 1973. The General Impossibility of Normative Accounting Standards. *The Accounting Review* 48 (4): 718–723.

Ding, K., X. Peng, and Y. Wang. 2019. A machine learning-based peer selection method with financial ratios. *Accounting Horizons* 33 (3): 75–87.

Durnev, A., R. Morck, and B. Yeung. 2004. Value-Enhancing Capital Budgeting and Firm-specific Stock Return Variation. *Journal of Finance* 59 (1): 65–105.

Ecker, F., J. Francis, P. Olsson, and K. Schipper. 2013. Estimation sample selection for discretionary accruals models. *Journal of Accounting and Economics* 56 (2–3): 190–211.

Engelberg, J., A. Ozoguz, and S. Wang. 2018. Know Thy Neighbor: Industry Clusters, Information Spillovers, and Market Efficiency. *Journal of Financial and Quantitative Analysis* 53 (5): 1937–1961.

Fama, E. F., and K. R. French. 1997. Industry costs of equity. *Journal of Financial Economics* 43 (2): 153–193.

Foucault, T., and L. Frésard. 2018. Corporate Strategy, Conformism, and the Stock Market. *The Review of Financial Studies* 32 (3): 905–950.

Franco, F., O. Urcan, and F. P. Vasvari. 2016. Corporate diversification and the cost of debt: The role of segment disclosures. *Accounting Review* 91 (4): 1139–1165.

De Franco, G., O.-K. Hope, and S. Larocque. 2015. Analysts' choice of peer companies. *Review of Accounting Studies* 20 (1): 82–109.

De Franco, G., S. P. Kothari, and R. S. Verdi. 2011. The benefits of financial statement comparability. *Journal of Accounting Research* 49 (4): 895–931.

Guay, W., D. Samuels, and D. Taylor. 2016. Guiding through the Fog: Financial statement complexity and voluntary disclosure. *Journal of Accounting and Economics* 62 (2–3): 234–269.

Guenther, D. A., and A. J. Rosman. 1994. Differences between COMPUSTAT and CRSP SIC codes and related effects on research. *Journal of Accounting and Economics* 18 (1): 115–128.

Hameed, A., R. Morck, J. Shen, and B. Yeung. 2015. Information, Analysts, and Stock Return Comovement. *Review of Financial Studies* 28 (11): 3153–3187.

Hoberg, G., and G. Phillips. 2010. Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis. *The Review of Financial Studies* 23 (10): 3773–3811.

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

———. 2018. Conglomerate Industry Choice and Product Language. *Management Science* 64 (8): 3469–3970.

Hoitash, R., and U. Hoitash. 2018. Measuring accounting reporting complexity with XBRL. *Accounting Review* 93 (1): 259–287.

Hrazdil, K., and T. Scott. 2013. The role of industry classification in estimating discretionary accruals. *Review of Quantitative Finance and Accounting* 40 (1): 15–39.

Hrazdil, K., K. Trottier, and R. Zhang. 2014. An intra- and inter-industry evaluation of three classification schemes common in capital market research. *Applied Economics* 46 (17): 2021–2033.

Jones, J. J. 1991. Earnings Management During Import Relief Investigations. *Journal of Accounting Research* 29 (2): 193.

Jones, K. L., G. V. Krishnan, and K. D. Melendrez. 2008. Do models of discretionary accruals detect actual cases of fraudulent and restated earnings? An empirical analysis. *Contemporary Accounting Research* 25 (1): 499–531.

Kadan, O., L. Madureira, R. Wang, and T. Zach. 2012. Analysts' industry expertise. *Journal of Accounting and Economics* 54 (2): 95–120.

Kahle, K. M., and R. A. Walkling. 1996. The impact of industry classifications on financial research. *Journal of Financial and Quantitative Analysis* 31 (03): 309–335.

Kelly, B., and A. Ljungqvist. 2012. Testing asymmetric-information asset pricing models. *Review of Financial Studies* 25 (5): 1366–1413.

Krishnan, J., and E. Press. 2003. The North American Industry Classification System and Its Implications for Accounting Research. *Contemporary Accounting Research* 20 (4): 685–717.

Larcker, D. F., S. A. Richardson, and I. Tuna. 2007. Corporate Governance, Accounting Outcomes, and Organizational Performance. *The Accounting Review* 82 (4): 963–1008.

Lee, C. M. C., P. Ma, and C. C. Y. Wang. 2015. Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics* 116 (2): 410–431.

Lehavy, R., F. Li, and K. Merkley. 2011. The effect of annual report readability on analyst following and the properties of their earnings forecasts. *Accounting Review* 86 (3): 1087–1115.

Lieberman, M. B., and S. Asaba. 2006. Why Do Firms Imitate Each Other? *Academy of Management Review* 31 (2): 366–385.

Loughran, T., and B. McDonald. 2017. The Use of EDGAR Filings by Investors. *Journal of Behavioral Finance* 18 (2): 231–248.

Luo, S., and N. J. Nagarajan. 2015. Information complementarities and supply chain analysts. *Accounting Review* 90 (5): 1995–2029.

Owens, E. L., J. S. Wu, and J. Zimmerman. 2017. Idiosyncratic Shocks to Firm Underlying Economics and Abnormal Accruals. *Accounting Review* 92 (2): 183–219.

Parrino, R. 1997. CEO turnover and outside succession: A cross-sectional analysis. *Journal of Financial Economics* 46 (2): 165–197.

Peterson, K., R. Schmardebeck, and T. J. Wilks. 2015. The Earnings Quality and Information Processing Effects of Accounting Consistency. *Accounting Review* 90 (6): 2483–2514.

Piotroski, J. D., and D. T. Roulstone. 2004. The Influence of Analysts, Institutional Investors, and Insiders on the Incorporation of Market, Industry, and Firm-Specific Information into Stock Prices. *The Accounting Review* 79 (4): 1119–1151.

Ramnath, S. 2002. Investor and Analyst Reactions to Earnings Announcements of Related Firms: An Empirical Analysis. *Journal of Accounting Research* 40 (5): 1351–1376.

Roll, R. 1988. R^2. *The Journal of Finance* 43 (3): 541–566.

Schmalensee, R. 1985. Do markets differ much? *American Economic Review* 75 (3): 341–351.

Shroff, N., R. S. Verdi, and B. P. Yost. 2017. When does the peer information environment matter? *Journal of Accounting and Economics* 64 (2): 183–214.

Thomas, S. 2002. Firm diversification and asymmetric information: Evidence from analysts' forecasts and earnings announcements. *Journal of Financial Economics* 64

(3): 373–396.

Zarzeski, M. T. 1996. Spontaneous harmonization effects of culture and market forces on accounting disclosure practices. *Accounting Horizons* 10 (1): 18–37.

Appendix A : Variable definitions

Variable		Definition
Main variables		
MISFIT	Misfit dummy	Indicator variable equal to one if the firm belongs to the misfit group for a given GICS6-year. See Section 2.3.1 for more details.
DIFF	Differentiation deciles	Ranked value (deciles from 1 to 10) of the firm-year distance calculated through the three-stage methodology. See Section 3.3.1 for more details.
CPX	Operational complexity	Firm-year measure of operational complexity based on industry relatedness. See Section 4.3.1 for more details.
Log_CPX		Log value of CPX, plus one.
Industry characteristics (all variable from Compustat)		
AVG_AT	Average total assets	Total assets (AT) + lagged total assets divided by 2
ROA	Return on assets	Income before extraordinary items (IB) divided by AVG_AT
CFO	Cash flow from operations	Operating activities net cash flow (OANCF) less extraordinary items and discontinued operations (XIDOC) divided by AVG_AT
CR	Current ratio	Current assets (ACT - CHE) on current liabilities (LCT - DLC)
AT_TURN	Asset turnover ratio	Sales (SALE) divided by AVG_AT
XSGA	Selling, general and administrative expense intensity	Selling, general and administrative expense (XSGA) divided by AVG_AT
COGS	Cost of goods sold intensity	Cost of goods sold (COGS) divided by AVG_AT
σ(CFO)	Operating cash flow volatility	Standard deviation of cash flow from operations (CFO) over the previous four years
σ(SALES)	Sales volatility	Standard deviation over the previous four years of sales (SALE) divided by AVG_AT
Size	Size	Natural logarithm of AVG_AT
BtoM	Book to market ratio	Book value of equity (CEQ) divided by market value of equity (CSHO*PRCC_F)
LEV	Leverage	Average long-term debt (DLTT) divided by AVG_AT
EVS	Enterprise Value-to-sales	Market value of equity (CSHO*PRCC_F) plus Debt in current liabilities (DLC) plus Long-term debt (DLTT) divided by total sales
EP	Earnings to price ratio	Income before extraordinary items (IB) divided by market value of equity (CSHO*PRCC_F)
Cap_int	Capital intensity	Property Plant and Equipment (Net) (PPENT) divided by AVG_AT

Variable		Definition
RD	Research and Development intensity	Research and Development expense (XRD) divided by total sales. Variable is set to 0, if XRD is missing.
PM	Profit margin	Income before extraordinary items (IB) divided by total sales
Accruals data		
TACC	Accruals	Earnings before extraordinary items from the cash flow statement (IBC), less operating cash flows (OANCF less XIDOC), scaled by average total assets (AVG_AT)
PPE	Property Plant and Equipment	Gross Property Plant and Equipment (PPEGT) divided by average total assets (AVG_AT)
DCF		Indicator variable equal to one if current year cash flows (OANCF) are negative
ΔReceivables	Changes in receivables	Change in receivables (RECT) divided by average total assets (AVG_AT)
ΔSales	Changes in sales	Change in sales (SALE) divided by average total assets (AVG_AT)
$\lvert DACC\rvert_{MODJ}$		Absolute abnormal accruals computed using the Modified Jones model (Dechow et al., 1995)
$\lvert DACC\rvert_{NL}$		Absolute abnormal accruals computed using the nonlinear model (Ball and Shivakumar, 2006)
$\lvert DACC\rvert_{DD}$		Absolute abnormal accruals computed using the model from Dechow and Dichev (2002)
$\lvert DACC\rvert_{MN}$		Absolute abnormal accruals computed using model from Dechow and Dichev (2002) modified by McNichols (2002)
Returns data (from CRSP)		
RET	Daily returns	Variable RET
INDRET	Industry daily returns	Value-weighted daily industry returns using all the firms' stock returns (RET) in the same GICS6 industry. For each firm *i,* the value weighted industry returns are calculated excluding firm *i.*
MVE	Market value of equity	Median of market value of equity (PRC*SHROUT) over the same period as Turnover
Turnover	Share turnover	Trade volume (PRC*VOL) divided by market value of equity (MVE). Median of monthly turnover calculated over months t-2 to t+10, where monthly turnover is the sum of daily trade volume divided by market value of equity (PRC*VOL/MVE)
σ(RET)	Returns volatility	Log value of the standard deviation of daily returns over the same period as Turnover
BidAsk	Bid-Ask spread	Median value of bid-ask spread (Ask-Bid) scaled by mid-point ((Ask+Bid)/2) over the same period as Turnover

Variable		Definition
Illiquidity	Amihud (2002) illiquidity measure	Median value of absolute value of returns divided by trade volume (PRC*VOL) over the same period as Turnover
Idio_Shock	Firm stock return-based idiosyncratic shock	For each firm *i,* I calculate a monthly value-weighted industry and market returns (excluding firm *i* stock return). Then, I regress firm *i*'s returns on the market and industry returns using two years of monthly data. I take the mean squared error of this regression as the Idio_Shock.
Peer_Idio_Shock		Mean value of *Idio_Shock* of all the other firms belonging to the same GICS6-Industry
Analyst data		
Accuracy	Analysts forecasts accuracy	Absolute value of the difference between the mean of the latest (1-year ahead annual) analysts' forecasts before the earnings announcements (MEANEST item in I/B/E/S) and the actual earnings (ACTUAL in I/B/E/S), multiplied by -100 and scaled by the firm share price at the end of the previous fiscal year
Coverage	Number of analysts following	Log value of the number of analysts following the firm (NUMEST item in I/B/E/S)
Dispersion	Analysts forecasts dispersion	Analyst forecast dispersion from the latest forecasts before the actual earnings announcement, multiplied by 100 and scaled by the firm share price at the end of the previous fiscal year
Misstatements data		
AAER	Accounting and Auditing Enforcement Releases	Data from Audit Analytics and AAER data from University of California, Berkeley's CFRM (Dechow et al. 2011) as extended bynd Bao et al. (2020) (available at https://github.com/JarFraud/FraudDetection)
RES	Restatement	Restatements data taken from AuditAnalytics (Non-Reliance Restatements file)
Control variables (all from Compustat except when specified)		
Volume	Trading volume	Log value of the trading volume (CSHTR_F) scaled by 1 million
Depreciation	Depreciation expense	Firm depreciation expense (DP) less industry median depreciation expense scaled by the firm's sales

Variable		Definition
Issue_3y	Debt or equity issuance	Indicator variable equal to one, if the firm has issued debt or equity (DLTIS>0) in the previous, current or future fiscal year, 0 otherwise (as in Peterson et al. 2015)
σ(ROA)	Standard deviation of Return on Assets	Standard deviation over four years of Returns on Assets (ROA)
SUE	Unexpected earnings	Absolute value of earnings surprise (IB less its lagged value), scaled by lagged price (PRCC_F)
NegSUE	Negative earnings surprise	Indicator variable equal to one, if the firm experienced a negative earnings surprise, 0 otherwise
LOSS		Indicator variable equal to one, i the firm experienced a loss (IB<0), 0 otherwise
NegSI	Negative special items	Absolute value of SPI item if SPI<0, 0 otherwise
Bog	Financial statement complexity	Based on the Bog Index in a Dataset provided by Brian P. Miller (https://kelley.iu.edu/bpm/activities/bogindex.html)
Days		Log value of the number of days between the last analyst forecast date (stat_pers) and the earnings announcement date (anndats_act) from I/B/E/S.
Op_shock		Indicator variable equal to one if the firm experienced at least one of the following operating shocks: industry change; large discontinued operations (Compustat DO greater than 5% of the sales); large merger or acquisition (Compustat AB>0); large restructuration (Compustat RCP greater than 5% of the sales); or large special items (Compustat SPI greater than 5% of the sales).
ChAuditor	Auditor change	Indicator variable equal to one if the firm changed its auditor (AU)
SP	Special items	Special items (SPI) scaled by average total assets (AVG_AT)
BigN		Indicator variable equal to one if the firm's auditor belongs to the Big N audit firms (0<AU<=8)
ΔInventory	Changes in inventory	Change in inventory (INVT) divided by average total assets (AVG_AT)
%_soft_assets	Percentage of soft assets	Total assets (AT) less cash and equivalents (CHE) and less Net Property, Plan and Equipment (PPENT), scaled by total assets (AT)
ΔCash_sales	Change in cash sales	Change in sales less change in receivables, scaled by lagged sales (SALE)
ΔROA	Change in profitability	Change in return on assets defined as income before extraordinary items (IB) divided by average total assets (AVG_AT)
NBS		Total number of business segments reported (from Compustat Segments). Segments without an attributed SIC code are excluded.

Variable	Definition
NBS_Sic2	Number of business segments at SIC2-level
Log_NBS	Natural logarithm value (plus one) of NBS

Appendix B: Excerpt of the GICS structure

Sector (GICS2)		Industry group (GICS4)		Industry (GICS6)		Sub-Industry (GICS8)	
10	**Energy**	**1010**	**Energy**	**101010**	**Energy Equipment & Services**	**10101010**	**Oil & Gas Drilling** Drilling contractors or owners of drilling rigs that contract their services for drilling wells
						10101020	**Oil & Gas Equipment & Services** Manufacturers of equipment, including drilling rigs and equipment, and providers of supplies and services to companies involved in the drilling, evaluation and completion of oil and gas wells.
				101020	**Oil, Gas & Consumable Fuels**	**10102010**	**Integrated Oil & Gas** Integrated oil companies engaged in the exploration & production of oil and gas, as well as at least one other significant activity in either refining, marketing and transportation, or chemicals.
						10102020	**Oil & Gas Exploration & Production** Companies engaged in the exploration and production of oil and gas not classified elsewhere.
						10102030	**Oil & Gas Refining & Marketing** Companies engaged in the refining and marketing of oil, gas and/or refined products not classified in the Integrated Oil & Gas or Independent Power Producers & Energy Traders Sub-Industries.
						10102040	**Oil & Gas Storage & Transportation** Companies engaged in the storage and/or transportation of oil, gas and/or refined products. Includes diversified midstream natural gas companies, oil and refined product pipelines, coal slurry pipelines and oil & gas shipping companies.
						10102050	**Coal & Consumable Fuels**

							Companies primarily involved in the production and mining of coal, related products and other consumable fuels related to the generation of energy. Excludes companies primarily producing gases classified in the Industrial Gases sub-industry and companies primarily mining for metallurgical (coking) coal used for steel production.
15	**Materials**	**1510**	**Materials**	**151010**	**Chemicals**	**15101010**	**Commodity Chemicals** Companies that primarily produce industrial chemicals and basic chemicals. Including but not limited to plastics, synthetic fibers, films, commodity-based paints & pigments, explosives and petrochemicals. Excludes chemical companies classified in the Diversified Chemicals, Fertilizers & Agricultural Chemicals, Industrial Gases, or Specialty Chemicals Sub-Industries.
						15101020	**Diversified Chemicals** Manufacturers of a diversified range of chemical products not classified in the Industrial Gases, Commodity Chemicals, Specialty Chemicals or Fertilizers & Agricultural Chemicals Sub-Industries.
						15101030	**Fertilizers & Agricultural Chemicals** Producers of fertilizers, pesticides, potash or other agriculture-related chemicals not classified elsewhere.
						15101040	**Industrial Gases** Manufacturers of industrial gases.
						15101050	**Specialty Chemicals** Companies that primarily produce high value-added chemicals used in the manufacture of a wide variety of products, including but not limited to fine chemicals, additives, advanced polymers, adhesives, sealants and specialty paints, pigments and coatings.

This Appendix presents an excerpt of the GICS structure from the latest available structure (effective after close of business September 28, 2018). The four level of the GICS structure are presented (from GICS2 to GICS8).

Appendix C: Calculation of DIFF for the fiscal year 2006 for the industry GICS 251010 “Auto Components”

Panel A: Logistic regression results

Variable	(1)
Intercept	0.6768
	0.49
Size	0.0108
	0.12
AT_TURN	0.5394
	0.32
BtoM	-1.7490**
	-2.5
CAP_INT	-0.3652
	-0.46
COGS	-1.0819
	-0.58
CR	0.3338**
	2.09
EP	-2.3268
	-1.54
EVS	-1.4585**
	-2.2
LEV	-1.7872
	-1.19
PM	1.7912
	1
RD	**8.8132*** **
	2.58
ROA	-2.7957
	-1.24
σ(CFO)	-0.3134
	-0.07
σ(SALES)	-1.2754
	-0.77
XSGA	**-7.1475*** **
	-3.39
Number of obs.	1,470

This model presents the results of a logistic regression for year 2006, using IND(j) as a dependent variable taking the value of 1 if the firm belongs to the GICS 251010 “Auto Components”.

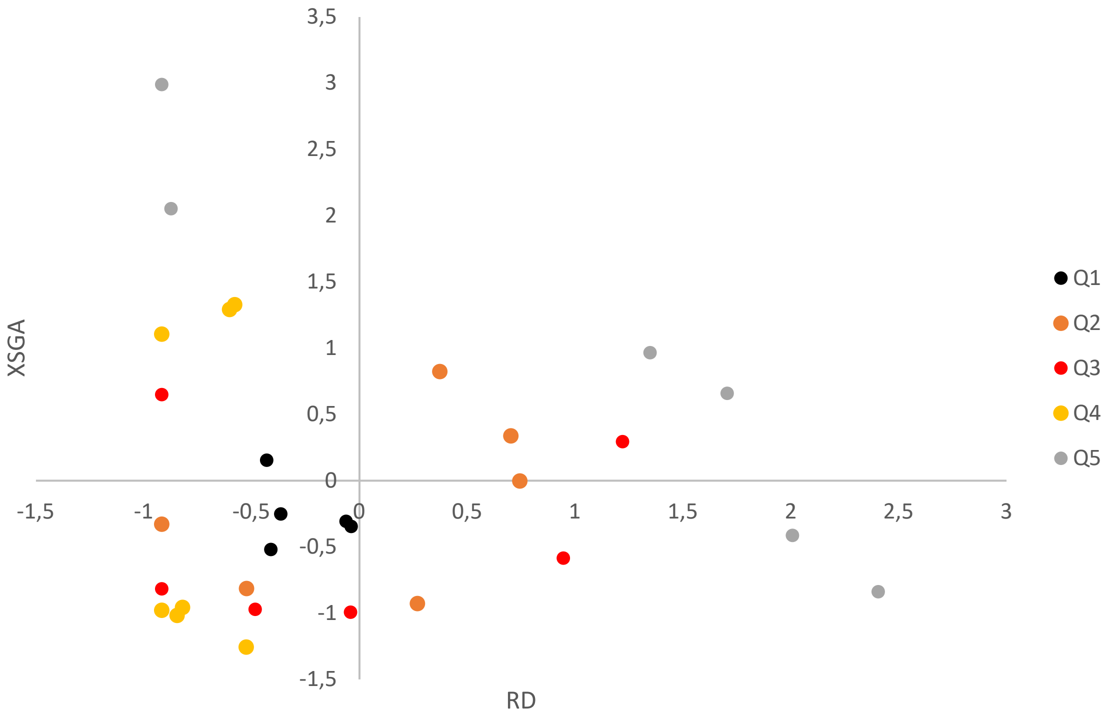
Panel B: Descriptive statistics

Variable: **XSGA**	N	Mean	StdDev	P10	Median	P90
Full sample	1440	0.30	0.23	0.06	0.25	0.61
GICS 251010 firms	30	0.18	0.12	0.07	0.15	0.34
Diff		-0.11***			-0.10***	
1 (Close to industry core)	3	0.15	0.01	0.14	0.15	0.15
2	2	0.16	0.06	0.12	0.16	0.20
3	3	0.23	0.05	0.18	0.22	0.28
4	3	0.10	0.04	0.07	0.09	0.14
5	4	0.13	0.09	0.07	0.09	0.26
6	2	0.15	0.09	0.09	0.15	0.22
7	4	0.06	0.02	0.04	0.07	0.07
8	3	0.33	0.01	0.31	0.34	0.34
9	4	0.28	0.12	0.13	0.28	0.42
10 (Far from industry core)	2	0.31	0.32	0.09	0.31	0.54
Diff (10-1)		0.16			0.16	

Variable: **RD**	N	Mean	StdDev	P10	Median	P90
Full sample	1440	0.04	0.12	0.00	0.00	0.10
GICS 251010 firms	30	0.02	0.02	0.00	0.01	0.05
Diff		-0.02			0.01***	
1 (Close to industry core)	3	0.02	0.00	0.01	0.02	0.02
2	2	0.01	0.00	0.01	0.01	0.01
3	3	0.03	0.00	0.03	0.04	0.04
4	3	0.01	0.01	0.00	0.01	0.03
5	4	0.02	0.02	0.00	0.01	0.04
6	2	0.02	0.03	0.00	0.02	0.05
7	4	0.00	0.00	0.00	0.00	0.01
8	3	0.00	0.00	0.00	0.01	0.01
9	4	0.04	0.03	0.00	0.05	0.06
10 (Far from industry core)	2	0.04	0.05	0.00	0.04	0.07
Diff (10-1)		0.02			0.02	

This table presents descriptive for the year 2006 for firms belonging to the GICS 251010. For two variables (XSGA, RD), the full sample and the industry mean (median) are presented, as well as means (medians) for each differentiation deciles. Mean (median) difference tests are t-test (Wilcoxon tests), where a *, ** and *** denote for differences significant at 1%, 5% and 10% respectively.

Panel C: Spatial representation



This graph presents the spatial distribution of all firms in the GICS 251010 "Auto Components" industry. All firms are presented on their two dimensions (RD, XSGA) representing the industry determinants in 2006. Industry centroid coordinates are (0,0). For simplification purpose, I present the data in quintiles rather than in deciles.

Appendix D: Calculation of *CPX* for Johnson & Johnson (gvkey=006266) and BASF (gvkey=017436) for 2014

Panel A: Calculation of the industry relatedness score for the industry sic28 "Chemicals and Allied Products" for the fiscal year 2014

	SIC2 code	Industry name	SIC2 code	Industry name	Euclidean distance
Complex industries	28	Chemicals and Allied Products	54	Food Stores	4.69
	28	Chemicals and Allied Products	13	Oil and Gas Extraction	3.70
	28	Chemicals and Allied Products	46	Pipelines, except Natural Gas	3.65
	28	Chemicals and Allied Products	55	Automotive Dealers & Service Stations	3.52
	28	Chemicals and Allied Products	58	Eating and Drinking Places	3.40
	28	Chemicals and Allied Products	53	General Merchandise Stores	3.40
	28	Chemicals and Allied Products	47	Transportations Services	3.30
	28	Chemicals and Allied Products	56	Apparel & Accessory Stores	3.28
	28	Chemicals and Allied Products	50	Wholesale Trade-Durable Goods	3.27
	28	Chemicals and Allied Products	33	Primary Metal Industries	3.26
	28	Chemicals and Allied Products	45	Transportation by Air	3.23
	28	Chemicals and Allied Products	59	Miscellaneous Retail	3.07
	28	Chemicals and Allied Products	51	Wholesale Trade-Nondurable Goods	3.06
	28	Chemicals and Allied Products	57	Furniture and Homefurnishings Stores	3.02
	28	Chemicals and Allied Products	14	Nonmetallic minerals, except fuels	2.94
	28	Chemicals and Allied Products	16	Heavy construction, except building	2.86
	28	Chemicals and Allied Products	79	Amusement and Recreation Services	2.82
	28	Chemicals and Allied Products	26	Paper and Allied Products	2.60
	28	Chemicals and Allied Products	30	Rubber and Miscellaneous Plastics Products	2.54
	28	Chemicals and Allied Products	82	Educational Services	2.48
	28	Chemicals and Allied Products	25	Furnitures and Fixtures	2.40
	28	Chemicals and Allied Products	39	Miscellaneous Manufacturing Industries	2.37
	28	Chemicals and Allied Products	20	Food and Kindred Products	2.34
Non-complex industries	28	Chemicals and Allied Products	37	Transportation Equipment	2.21
	28	Chemicals and Allied Products	23	Apparel and Other Textile Products	2.19
	28	Chemicals and Allied Products	49	Electric, Gas and Sanitary Services	2.16
	28	Chemicals and Allied Products	34	Fabricated Metal Products, except Machinery	2.12
	28	Chemicals and Allied Products	27	Printing, Publishing, and Allied Industries	2.11
	28	Chemicals and Allied Products	80	Health Services	2.08
	28	Chemicals and Allied Products	35	Industrial Machinery and Equipment	2.00
	28	Chemicals and Allied Products	48	Communications	1.82
	28	Chemicals and Allied Products	73	Business Services	1.78
	28	Chemicals and Allied Products	87	Engineering and Management Services	1.77
	28	Chemicals and Allied Products	36	Electronic and Other Electric Equipment	1.28
	28	Chemicals and Allied Products	38	Instruments and Related Products	0.66

This panel presents the distance between the industry *sic28 "Chemicals and Allied Products"* and every other industries for the fiscal year 2014. The median distance for the fiscal year 2014 is 2.34. Industries above (below) the double line border represents complex (non-complex) industries for firms belonging to the *sic28* industry.

Panel B: Calculation of CPX for Johnson & Johnson (gvkey=006266) and BASF (gvkey=017436)

Company name	**GVKEY**	**Industry Membership**	**Business segment (SIC2-level)**	**Industry relatedness score**	**Median score for 2014**	**Complex business segment?**
JOHNSON & JOHNSON	006266	*Sic28 "Chemicals and Allied Products"*	*Sic28 "Chemicals and Allied Products"*	0.00	2.34	No
JOHNSON & JOHNSON	006266	*Sic28 "Chemicals and Allied Products"*	*Sic38 "Instruments and Related Products"*	0.66	2.34	No
BASF	017436	*Sic28 "Chemicals and Allied Products"*	*Sic28 "Chemicals and Allied Products"*	0.00	2.34	No
BASF	017436	*Sic28 "Chemicals and Allied Products"*	*Sic13 "Oil and Gas Extraction"*	3.70	2.34	Yes