Differentiation in Microenterprises

Natalie A. Carlson *1

1The Wharton School, 3620 Locust Walk, Philadelphia, PA 19104, USA

Abstract

Small unregistered firms contribute to a substantial proportion of global economic activity, particularly in developing regions. In explaining variation in productivity in these types of informal firms, research has focused primarily on the adoption of effective business practices and access to capital, with little focus on fundamental positioning. This article explores the nature of differentiation in microenterprises, introducing a text-based measure of differentiation using state-of-the-art sentence embeddings. Using a combined sample of nearly 10,000 microenterprises across eight developing countries, I estimate that a standard deviation increase in differentiation is associated with approximately an 11 percent increase in revenues and an eight percent increase in profit, relative to the sample mean. I show that the relationship between differentiation and performance is substantially stronger for male than for female microenterprise owners, and that the propensity to differentiate increases with years of education and decreases with age. Finally, I estimate the impact of common policy interventions on microenterprise differentiation. The results suggest that standard business skills training interventions have little effect on differentiation and that access to individual-liability microfinance may actually decrease it.

Keywords: Microenterprise, Emerging Markets, NLP, Sentence Embeddings, Competitive Strategy

*Please send correspondence to ncarlson@wharton.upenn.edu
1 Introduction

An enormous percentage of global economic output is produced by informal microenterprises, particularly in developing economies (Medina and Schneider, 2017). Such firms are typically unregistered, hire few employees outside the immediate family, and operate within a harsh institutional environment (Mair and Marti, 2009; Khanna and Palepu, 1997; Sutter et al., 2019). In understanding what factors drive the performance of such small firms, the primary focus in much research and policy has been on incremental improvements in productivity through encouraging the adoption of better business practices or facilitating access to capital (e.g., McKenzie and Woodruff, 2013). By contrast, most policy interventions have focused little on fundamental business positioning. Given that informal firms are often formed through a process of imitating existing firms and compete on a cost basis (Alvarez and Barney, 2014; Block et al., 2015), however, the ability to provide a differentiated offering may be an important predictor of performance.

Due to the difficult conditions under which many microenterprises operate, it is broadly assumed differentiation will be difficult to achieve, as proprietors cope with constraints on financial and cognitive resources (Doering, 2016; Block et al., 2015; Bruton et al., 2013). This article aims to examine to what extent microenterprises are able to differentiate – and how – taking the approach of a question-driven, hypothesis-free empirical analysis (Graebner et al., 2022). The goal of the empirical exercise is to shed light on which constraints are most crucial in limiting microenterprises’ ability to differentiate, and whether or not these constraints can be released. The results provide clarity on the shape of the relationship between differentiation and performance in microenterprise, which types of microentrepreneurs are most likely to differentiate, and what impacts existing policy interventions have on differentiation outcomes.

The empirical analysis examines a combined dataset created from survey data in eight prior
studies of microenterprises. Identifying open-access studies in which survey participants were asked to describe the activities of their business using open-ended text, I used the descriptions to create a text-based measures of differentiation using sentence embeddings based on the BERT network (Reimers and Gurevych, 2020b). The combined sample comprises approximately 10,000 microenterprises across eight countries, with data in five different languages. The analysis estimates that a standard deviation increase in differentiation is associated with approximately nine percent more employment, 11 percent higher revenues, and eight percent higher profits relative to the mean. This relationship is evident within retail, manufacturing, and agriculture firms, though not within service firms.

Further analysis examines which types of microenterprise owners are most likely to differentiate, providing some clues to one of the motivating questions of the study: are the most important constraints on differentiation in microenterprises due to limitations in human capital, financial resources, or some other factor? I find that propensity to differentiate decreases with age in a strikingly linear relationship. Differentiation also increases with years of education, with a particular boost after secondary school. Male microenterprise owners have on average more differentiated businesses than female owners, and the relationships between differentiation and performance measures are approximately six to nine times stronger for men than they are for women. The heterogeneity suggests that both cognitive and resource constraints may be at play in limiting microenterprises’ ability to differentiate.

A natural follow-up inquiry is how best such constraints to differentiation may be released. Leveraging the fact that seven of the eight studies included in the sample involve a randomized experiment, I examine whether any of the individual interventions had treatment effects on the differentiation measure. The interventions in the individual studies fall into two broad categories, training programs and microloan access, which could theoretically function to loosen human-capital- and resource-based constraints, respectively. I find that the training programs had consistent null effects on the differentiation measure, and that two
out of the three individual-liability microloan interventions appear to have had a negative impact on differentiation. I discuss the implications of these findings and some suggestions for interventions that might prove more promising.

This article offers contributions empirical, methodological, and practical. Empirically, the analysis provides to my knowledge the first in-depth descriptive analysis of differentiation in microenterprise and its relationship with performance. The results suggest a robust positive relationship, though not one that is uniform across all microenterprises: the association is considerably weaker in female-run businesses, for example, and virtually absent in service businesses. By examining differentiation in very small firms, in which the firm is effectively synonymous with the proprietor, the analysis also provides many interesting insights at the individual level with the potential to inform future theory – considering, for example, how age, gender, and formal education impact the ability to stake out a differentiated market position.

Methodologically, this work demonstrates the usefulness of text-based measures in capturing the positioning of one firm vis-à-vis a set of related firms (e.g., Guzman and Li, 2021; Hoberg and Phillips, 2010, 2016), and in particular demonstrate the value of sentence embeddings (Reimers and Gurevych, 2020b). In introducing a differentiation measure using BERT-derived sentence embeddings, this article shows how contextual embeddings may capture relatedness in business concepts that could be missed by more naive approaches to measuring text distance. The supplemental materials provide code and instruction for replicating the measures described here, creating a roadmap for future researchers.

From a practitioner standpoint, this work aims to demonstrate the relevance of strategic positioning in microenterprises. While strategy and management researchers have posited that differentiation could be an effective source of competitive advantage in this population of businesses (e.g., Block et al., 2015; Bruton et al., 2011), these constructs are not salient in most policy-adjacent work. Indeed, the words “differentiation”, “strategy”, or “positioning”
cannot be found in any of several excellent research summaries of policy programs directed at microenterprises (e.g., McKenzie and Woodruff, 2013; McKenzie, 2020; Quinn and Woodruff, 2019; Jayachandran, 2021). While this analysis does not demonstrate a causal relationship, the robust association between differentiation and performance described in the results warrants an examination of a) whether any cost-effective interventions can successfully help microentrepreneurs achieve a more differentiated offering, and b) if such differentiation then translates causally to improved performance. The results of this analysis provide suggestive evidence as to what types of interventions may prove impactful and which population of microentrepreneurs might be most effective to target.

This article is organized in six sections: the following section will discuss theory on the relevance of differentiation in microenterprises, and the usefulness of text-based approaches in measuring strategy constructs. The Data and Methods section will describe the collection of the component datasets and provide detail on the text analysis methods. The Results section will discuss the relationship between the differentiation measure and business characteristics, and the Discussion and Conclusion sections will explore implications of the findings and future research opportunities.

2 Background and Theory

2.1 Microenterprises and the Informal Economy

The informal economy has been estimated to produce roughly a third of economic output in Asia and South America, as well as over 40 percent of output in sub-Saharan Africa (Medina and Schneider, 2017). The majority of this informal economic activity is generated by microenterprises with few or no employees (International Labour Organization, 2019), operating without formal registration (Bruton et al., 2012; Webb et al., 2009; Assenova and Sorenson, 2017). While these small businesses and their proprietors are often considered the backbone of what makes up “the bottom of the pyramid” (Prahalad, 2009; Seelos and Mair,
and “the last great frontier in international business” (Bruton et al., 2011), they have not received substantial attention in management and innovation scholarship, despite calls for more research focused on the world’s poor (Bruton, 2010; George et al., 2012).

The potential for unlocking economic growth by focusing policy on small informal firms is hotly debated. LaPorta and Schleifer (2008) criticized what they called the “romantic” view of informal firms, epitomized by de Soto (1989; 2000), which suggests that informal microenterprises could be engines of growth if certain barriers, such as cumbersome regulations or financial constraints, were to be removed. Instead, they suggested that informal firms are persistently inefficient and unproductive, and are apt to disappear as larger, more productive formal firms enter a market. Schoar (2010) has similarly suggested that subsistence entrepreneurs are categorically different from “transformational” (i.e., high-growth) entrepreneurs, and that the vast majority of microenterprises are not likely to grow or generate significant employment.

A number of global policy interventions aimed at microenterprises have been tested by researchers working with NGOs and governments (see McKenzie and Woodruff, 2013). Many of these interventions involve a training component, intended to increase microenterprise proprietors’ human capital and, in particular, encourage adoption of more formalized and efficient business practices (e.g., Bruhn and Zia, 2013; Valdivia, 2015). Such training programs have had uneven impacts on firm performance, though a recent synthesis of the literature argues that most studies are underpowered to observe effects (McKenzie, 2020). Another common category of policy intervention involves improving microenterprises’ access to capital, either through pure cash grants (e.g., Blattman et al., 2016; Fafchamps et al., 2014) or through improving access to credit, often microfinance (e.g., Banerjee et al., 2015; Attañasio et al., 2015). Randomized interventions involving the provision of microfinance have generally had disappointing effects on firm outcomes, though some studies providing cash grants have found large and positive impacts (Jayachandran, 2020).
2.2 Competitive Strategy in Microenterprises

Small-scale informal entrepreneurs tend to compete by lowering their costs rather than by attempting to differentiate (Block et al., 2015). Often their businesses are formed through a “replication” process of imitating existing operations (Alvarez and Barney, 2014; Dencker et al., 2021). In developing economies in which microenterprises make up a large percentage of economic activity, it is therefore not uncommon to see nearly identical small businesses side by side in the same marketplace (Delecourt and Ng, 2019). Prior work suggests that considerable limitations may constrain microenterprises’ ability to pursue differentiation (Doering, 2016). These limitations fall broadly into two categories: limitations in financial resources, and cognitive or human capital explanations.

Developing a differentiated offering is likely to take more time and start-up capital than starting an imitation business: entrepreneurs may have to source differentiated inputs, develop a new production process, or engage in trial and error. Microenterprise owners in developing countries typically operate under severe financial constraints, however, with scant savings and limited access to credit (Alvarez and Barney, 2014). These financial limitations may make pursuing differentiation less tenable, particularly if the potential returns are higher but more uncertain (Dencker and Gruber, 2015), or likely to be preceded by a long delay, lowering the net present value (Seelos and Mair, 2007). The risk associated with a delay in cash flow is amplified in environments with little social safety net and the potential for dramatic economic or political change (Dimitriadis, 2021). This institutional uncertainty may reduce the value of long-term strategic planning (Hiatt and Sine, 2014) and drive entrepreneurs to favor a strategy with immediate and reliable returns. Access to inexpensive labor through family members may also lead microentrepreneurs to favor a low cost strategy over differentiation (Block et al., 2015).

The financial constraints faced by many microenterprise owners is compounded by, and acts upon, the significant cognitive burden of operating under conditions of poverty. The search
for a novel or differentiated business opportunity can be cognitively as well as financially taxing (Brixiova et al., 2015), and conditions of scarcity directly impede cognitive function (Mani et al., 2013; Anuj K. Shah et al., 2012). Only once basic needs are met can entrepreneurs engage in experimentation or trial and error (Dencker et al., 2021) and develop a future orientation, crucial for microenterprise performance (Bruton et al., 2011). At a baseline level, a lack of formal education may also make it more difficult for microentrepreneurs to execute on a differentiation strategy (Block et al., 2015; Alvarez and Barney, 2014).

Should microenterprise owners overcome these constraints and achieve a differentiated offering, therefore, theory suggests they should observe considerable performance benefits to doing so. Very little work demonstrates this relationship empirically: in one exception, a study of 201 small business owners in Kenya found that differentiation-related innovation was associated with positive business performance (Bradley et al., 2012). Understanding more comprehensively the shape of the relationship between differentiation and performance – as well as which types of microenterprise owners are able to successfully differentiate, and whether any policy interventions directly impact differentiation – may shed light on whether this aspect of competitive strategy is a potential lever for improving the livelihoods of the world’s working poor.

### 2.3 Text-Based Measures in Strategy Research

There exists a growing tradition of computational approaches to text in the social sciences (Teodorescu, 2017; Grimmer and Stewart, 2013; Gentzkow et al., 2019). In studies of strategic management, computational text analysis has often been used to characterize softer constructs such as communication: for example, Choudhury et al. (2019) and Pan et al. (2018) use text from unstructured interviews and earnings call transcripts, respectively, to measure CEO communication.

One of the richest applications of text analysis in strategy research has been in capturing
patent similarity or scope (e.g., Arts et al., 2018; Kuhn and Teodorescu, 2021; Kuhn and Thompson, 2019; Wang et al., 2019). Studies in this area frequently use distance measures, such as cosine distance or the Jaccard index, to map the distance between patent applications. Patent text may be converted to vector space using a bag-of-words approach\(^1\) (e.g., Younge and Kuhn, 2016; DeGrazia et al., 2021) or by using word embeddings, which may be domain-specific (as in Risch and Krestel (2019)) or employ pre-trained models (as in Whalen et al. (2020), using Doc2Vec). Relatedly, Bowen et al. (2019) use cosine similarity measures to measure novelty in patent text, while Kaplan and Vakili (2015) employ topic modeling to identify the emergence of novel topics in patents.

Other work uses text-based distance measures to capture constructs related to firms’ positioning or product diversification. For example, Hoberg and Phillips use measures of text distance to demonstrate that transactions are more likely between firms that use similar product market language (2010), that firm R&D and advertising are associated with subsequent differentiation from competitors (2016), and that language overlap can illustrate synergies in conglomerate product lines (2017). Barlow et al. (2019) undertake a similar approach in capturing the positioning of various offerings on the Google Play platform. Choi et al. (2021) use a corpus of 10-Ks to create text-based measures of diversification and examine its relationship to performance. Vicinanza et al. (2020) use BERT embeddings to identify prescient ideas in earnings call transcripts, allowing them to recognize visionary thinkers. Most similarly, Guzman and Li (2021) use a cosine distance measure with word embeddings to capture differentiation in the founding strategy of entrepreneurial firms, defining the average distance from the closest five incumbents. They demonstrate that this measure of differentiation predicts equity growth events such as IPO or acquisition, as well as financing performance in the early stages.

\(^1\)Typically tf-idf weighted, a way of adjusting for overall word frequency.
3 Data and Methods

3.1 Data

3.1.1 Selecting Studies for the Combined Analysis

To select studies for the purpose of the combined analysis, I searched for studies of microenterprises with open-access datasets. Fortunately, there have been a number of relevant studies conducted by development economists over the past two decades, along with a strong norm of sharing replication data. I searched three databases for relevant studies: the Harvard Dataverse’s Datahub for Field Experiments in Economics and Public Policy, the World Bank’s Microdata Library, and the openICPSR (Inter-university Consortium for Political and Social Research). Criteria for inclusion were straightforward: the study needed to have a text variable describing the microenterprise and at least one of the relevant performance measures of interest (employment, revenues, or profits). I found that while a number of studies had collected these open-ended text descriptions, in many cases they had not been used. The studies included in the combined analysis are described in Table 1.
<table>
<thead>
<tr>
<th>Label</th>
<th>Associated Reference</th>
<th>Sample Size</th>
<th>Year Collected</th>
<th>Country</th>
<th>Intervention Type</th>
<th>Targeted Population</th>
<th>Language of Text Var</th>
<th>Degree of Informality</th>
<th>Dataset Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mongolia 2008</td>
<td>Attanasio et al. (2015)</td>
<td>360</td>
<td>2008</td>
<td>Mongolia</td>
<td>Joint- or individual-liability microcredit programs</td>
<td>Rural women in Mongolian villages</td>
<td>English</td>
<td>Article does not specify enterprises’ registration status; a 2015 survey found that 68 percent of Mongolia businesses surveyed were formally registered (International Labour Organization, 2015)</td>
<td>Attanasio et al. (2018)</td>
</tr>
<tr>
<td>Philippines 2008</td>
<td>Karlan and Valdivia (2011)</td>
<td>1214</td>
<td>2007</td>
<td>Philippines</td>
<td>Access to credit-scoring and microloans</td>
<td>Microentrepreneurs applying for first-time loans with marginal creditworthiness</td>
<td>English/Filipino</td>
<td>Nearly all businesses in the sample reported either a barangay business license or mayor’s permit</td>
<td>Karlan and Zinman (2016)</td>
</tr>
<tr>
<td>Uganda 2007</td>
<td>Blattman et al. (2016)</td>
<td>843</td>
<td>2007</td>
<td>Uganda</td>
<td>$150 cash, training, and business supervision</td>
<td>Ultra-poor young women in post-war Northern Uganda</td>
<td>English/Acholi</td>
<td>Article does not specify enterprises’ registration status; informal sector accounts for 75 percent of employment in Uganda (Balloon Ventures, 2018)</td>
<td>Annan et al. (2015)</td>
</tr>
<tr>
<td>Togo 2013</td>
<td>Campos et al. (2017)</td>
<td>1499</td>
<td>2013</td>
<td>Togo</td>
<td>Business training or personal initiative training</td>
<td>Microentrepreneurs applying to a World Bank-financed government program</td>
<td>French</td>
<td>All businesses were unregistered as a selection criterion</td>
<td>McKenzie (2021b)</td>
</tr>
<tr>
<td>Bangladesh 2010</td>
<td>McKenzie (2010)</td>
<td>1601</td>
<td>2010</td>
<td>Bangladesh</td>
<td>None</td>
<td>Randomly selected business owners in urban Bangladesh</td>
<td>English</td>
<td>All businesses were unregistered as a selection criterion</td>
<td>McKenzie (2021a)</td>
</tr>
<tr>
<td>Mexico 2005</td>
<td>McKenzie and Woodruff (2008)</td>
<td>220</td>
<td>2005</td>
<td>Mexico</td>
<td>Shocks to capital stock in the form of cash or equipment</td>
<td>Microenterprise owners in Leon, Mexico, surveyed in the Mexican National Survey of Microenterprises</td>
<td>Spanish</td>
<td>Roughly 66 percent of surveyed businesses were unregistered</td>
<td>McKenzie and Woodruff (2021)</td>
</tr>
<tr>
<td>Kenya 2013</td>
<td>McKenzie and Puerto (2021a)</td>
<td>3521</td>
<td>2013</td>
<td>Kenya</td>
<td>Business skills training</td>
<td>Female microenterprise owners in markets in Kenya under age 55</td>
<td>English</td>
<td>Roughly 55 percent of businesses in the sample were unregistered</td>
<td>McKenzie and Puerto (2021b)</td>
</tr>
<tr>
<td>Zimbabwe 2017</td>
<td>Carlson and Hager (2021)</td>
<td>623</td>
<td>2017</td>
<td>Zimbabwe</td>
<td>Five days of business skills training</td>
<td>Microenterprise owners age 18-35 applying to a training program through the International Youth Foundation</td>
<td>English</td>
<td>Article does not specify enterprises’ registration status; Zimbabwe has the second largest informal sector in the world and most small businesses are unregistered (Medina and Schneider, 2018)</td>
<td>McKenzie and Puerto (2021b)</td>
</tr>
</tbody>
</table>

Sample size describes the number of observations with business text description.
3.1.2 Text Variable

One of the key criteria for dataset inclusion was the presence of a text variable describing the key activities of each microenterprise. Table 2 summarizes some common concerns and questions that might arise when determining if a text variable is appropriate for the sentence embedding method used in this article. Sentence embeddings are intended for vectorizing the semantic meaning of short texts, typically a paragraph at most (BERT and related models tend to be limited to 512 tokens). Multilingual text does not typically present a problem – transformer models have been trained across several dozen of the world’s most common languages – but texts that mix languages within sentences (e.g., “Selling camisas and zapatos”) will lose meaning in the conversion to vector embeddings. Likewise, typos and misspelled words will cause the embeddings to lose semantic meaning. Text variables that are not technically an open-ended response but rather a categorical variable with many detailed text labels may also function for the techniques discussed in this article, given sufficient variation.

**Table 2: Relevance of Common Text Data Concerns for Sentence Embeddings**

<table>
<thead>
<tr>
<th>Issue</th>
<th>Level of Concern</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texts are multilingual</td>
<td>Low</td>
<td>Several of the SBERT transformer models are multilingual and can compute semantic similarity across 50+ languages (mixing languages within observation is more of a concern).</td>
</tr>
<tr>
<td>Texts are (very) short</td>
<td>Low</td>
<td>Meaningful embeddings can be encoded from texts as short as one or two words (e.g., “barber”), depending on the purpose of the analysis.</td>
</tr>
<tr>
<td>Texts are (very) long</td>
<td>High</td>
<td>BERT-derived models are typically limited to 512 tokens and not intended for longer texts.</td>
</tr>
<tr>
<td>Texts are messy</td>
<td>Medium</td>
<td>Typos and misspellings introduce noise and reduce usefulness in deriving semantic meaning from the sentence embeddings. Note that BERT-derived embeddings, being based on natural language texts, are somewhat robust to typos.</td>
</tr>
<tr>
<td>Texts are (too) clean</td>
<td>Medium</td>
<td>Texts that are pre-cleaned or pre-categorized may still be useful, given sufficient variation (e.g., the methods described here could be used on a categorical variable with 10+ text labels).</td>
</tr>
</tbody>
</table>

As an example of how several issues mentioned in Table 2 were handled in the component
datasets, the text variable in Blattman et al. (2016) was uncleaned, containing many typos and written in a mix of English and Acholi, a relatively uncommon language used in Northern Uganda (not one of the 50+ languages covered in the pre-trained model). To reduce the noise introduced by these issues, I manually cleaned the texts, translating the Acholi words to English using an online dictionary. The differentiation measure based on the “clean” version of the text variable had a correlation coefficient of 0.61 with its counterpart based on the “messy” version – not significant enough to materially change any results, but nevertheless a prudent exercise. BERT embeddings and embeddings from related models are somewhat robust to typos, however, as they are trained on natural language corpuses that include typos as a matter of course – an additional advantage over bag-of-words-based measures. For example, the sentence-BERT embeddings for the two texts “charcoal” and “charcaol” have a cosine distance of 0.5405, while any term-count-based approach would naturally give the two texts a maximum distance value of 1.0.

3.1.3 Other Measures

When available, I captured measures of the microenterprises’ size and performance – total employment, monthly revenues, and monthly profits\(^2\). Within dataset, I created a mean-adjusted version of each performance indicator for comparability. I also captured data on the individual characteristics of the primary owner of the microenterprise from each dataset: age in years, gender, and total years of formal schooling. Finally, I captured the primary sector of the microenterprise: retail, services, manufacturing, agriculture, or other. In most of the datasets, this sector information was available from a pre-specified sector variable; in several datasets (Mexico, Bangladesh, Togo, Mongolia, and Uganda) I determined the appropriate sector by hand.

As most of the studies involved some kind of randomized intervention, I used baseline (pre-  
\(^2\)In the Uganda, and Togo datasets, employment data were either not available or missing for the majority of businesses; in the Mexico sample, profit data were not available; in the Mongolia sample, only revenue data were available. See Table 1 for more details.

13
treatment) data when available for the purpose of the primary analyses. In the Philippines, Uganda, and Zimbabwe samples only post-treatment data were available, however, and in those cases I controlled for treatment assignment in all analyses. Separately, I collected post-treatment measures for the purposes of determining whether the individual treatments had any causal impacts on the differentiation measure (see Results section 4.4), as well as indicators of each individual’s treatment assignment status.

3.2 Differentiation Measure

Table 3 summarizes the procedure for creating the Differentiation Measure from the text variable. In the first step, the text description of the business activities is converted to a vector representation using a pre-trained sentence embedding model. Embeddings are vector representations of terms derived from co-occurrence of words in a body of text. After being trained on a large text corpus, word embedding vectors can capture semantic relationships between terms, most canonically in the sense of analogies: for example, the vector equation “King - Man + Woman” would produce the vector for “Queen” (Mikolov et al., 2013). While popular word embedding algorithms such as Word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) have a single vector representation for any given term, context-dependent embedding approaches such as Google’s BERT (Devlin et al., 2019) take the surrounding context into account: the term “bank” would have a different representation in the phrase “bank teller” than in the phrase “the left bank”, for example.

In this article I employ the sentence-BERT technique (Reimers and Gurevych, 2020b) using the sentence-transformers package in python. Sentence-BERT is an efficient modification of the BERT network that provides multiple pre-trained models to compute meaning semantic representations of sentences in over 100 languages (Reimers and Gurevych, 2020a). The result is that each text description is converted to a vector of length 768: a representation that appears meaningless to the human eye but in fact encodes significant semantic meaning.

3The specific pretrained model used here is paraphrase-multilingual-mpnet-base-v2.
This meaning becomes clear when the vector representations are directly compared to one another, revealing the distance or closeness between embeddings.

**Table 3: Creating the Differentiation Measure**

<table>
<thead>
<tr>
<th>within each dataset ( k ):</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Encode sentence embeddings, converting each text to vector ( \text{SBERT models are typically length 768} )</td>
</tr>
<tr>
<td>“buy and sell phones and electronics” ( \rightarrow [-0.05013 \ -0.01732 \ ... \ -0.00491] )</td>
</tr>
<tr>
<td>2. Compute pairwise cosine distance between embedding representations for each pair of businesses ( a, b ) within ( k )</td>
</tr>
</tbody>
</table>
| \[
1 - \cos(a, b) = (1 - \frac{a \cdot b}{\|a\|\|b\|}) = (1 - \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2 \sqrt{\sum_{i=1}^{n} (b_i)^2}}})
\] |
| 3. For each firm \( a \) in \( k \), take the mean pairwise cosine distance between itself and every other firm |
| \[
\sum_{b=1}^{n} 1 - \cos(a, b) \]
| \[
\frac{n}{n}
\] |
| 4. The resulting value is the **Differentiation Score**, a measure of how distant the focal business is from its peers within the embedding vector space. Standardize with mean zero and standard deviation of one for comparability across datasets. |

The second step, therefore, calculates the pairwise cosine distance between the vector representations of each pair of businesses within the dataset\(^4\). For each business, the mean pairwise cosine distance between itself and each other business in the dataset can then be calculated. The **Differentiation Measure**, is the standardized value of this average cosine distance. The measure effectively captures how close a business is to its peer businesses within the vector embedding space; a low differentiation score indicates that its semantic meaning is close to many of the other businesses in the dataset, while a high score indicates that it is quite distant.

Table 4 displays example texts representing the least and most differentiated businesses within each sample, as well as those businesses around the median value of differentiation. To use the data from Carlson and Hager (2021) as an example, microenterprises with especially low scores tend to refer to nonspecific retail: “buying and selling”, “rearing and selling”,

\(^4\)Cosine distance is a commonly used distance metric for purposes similar to the one in this article, but by no means the only relevant distance measure; the results in this analysis are robust to using the Jaccard Index as well.
and “selling second hand goods” were some responses on this end of the spectrum. Those with the highest differentiation scores tend to offer more specialized services or products: “coffin maker”, for example, or “dance training for school students.” Responses also capture vertical differentiation to a limited degree: one of the highly differentiated Kenyan businesses reports selling “designer” clothes, for example, while a business in the Philippines sample describes selling “medicines at low price.”

**Table 4: Examples of Most and Least Differentiated Businesses Within Samples**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Lowest Scores</th>
<th>Median Scores</th>
<th>Highest Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mongolia 2008</td>
<td>“Sewing”</td>
<td>“Felt cashmere leather production”</td>
<td>“Cafe restaurant”</td>
</tr>
<tr>
<td>Philippines 2008</td>
<td>“Sari-sari store items”; “Sells soft drinks chips etc”</td>
<td>“Sell and deliver peanut butter”; “Sells medicines at low price, food supplements”</td>
<td>“Lights and sounds for events programs”; “Computer rental”; “Provide services, office uniform, school uniform”</td>
</tr>
<tr>
<td>Uganda 2007</td>
<td>“Cotton and charcoal”; “Tobacco”</td>
<td>“Firewood dry grass”; “Selling malt liquor”</td>
<td>“Drugs for animals”; “Selling spare parts for bikes”</td>
</tr>
<tr>
<td>Togo 2013</td>
<td>“Vente de nourriture”; “Vente des oeufs”</td>
<td>“Tapisserie”; “Vente de poissons fumés”</td>
<td>“Reportage photovideo”; “Reparation des appareils electroniques”</td>
</tr>
<tr>
<td>Bangladesh 2010</td>
<td>“Wholesale of textiles clothing and footwear”; “Other retail sale in specialized stores”</td>
<td>“Manufacture of knitted and crocheted fabrics and articles”; “Wholesale of electronic and telecommunications parts and equipment”</td>
<td>“Cutting shaping and finishing of stone”; “Library and archives activities”</td>
</tr>
<tr>
<td>Mexico 2005</td>
<td>“Venta de jugos”; “Venta de dulces y fruta”</td>
<td>“Muñequitos de peluche”; “Venta y producción de zapato”</td>
<td>“Venta de equip de telefonos celulares”; “Perfumería”</td>
</tr>
<tr>
<td>Zimbabwe 2017</td>
<td>“Rearing and selling”; “Selling second hand goods”</td>
<td>“Carpentry”; “Selling dishwashers and toilet cleaners”</td>
<td>“Coffin maker”; “Secretarial consultancy”; “Dance training for school students”</td>
</tr>
</tbody>
</table>

3.3 Why Use Sentence Embeddings?

As shown in some of the prior work measuring patent similarity, it is possible to create a measure similar to the Differentiation Score without going to the trouble of using the sentence
embeddings model to compute vector representations. Related similarity measures have been computed using bag-of-words representations of texts, in which the vector representations measure word counts where each column represents a unique term in the vocabulary. One of the limitations of this term vector approach is that two very similar businesses could use slightly different language to describe their operations. For example, under the naive approach described in the prior section, the semantic similarity between “grocery store” and “food seller” would not be captured because the two phrases share no terms in common. Additionally, the bag-of-words approach could capture the extent to which the business proprietor uses more sophisticated or precise language – a construct plausibly correlated with ability – as opposed to the underlying uniqueness of the business proposition.

To illustrate the value of using the sentence embeddings approach, Table 5 displays calculations of cosine distance between sample pairs of business with high degrees of semantic overlap. The first column calculates the cosine distance between the bag-of-words vector representations, while the second column calculates cosine distance between the sentence embedding vector representations. For example, the businesses “sell chickens” and “poultry” are substantively highly related, but the bag-of-words cosine distance measure between the two is 1.0 (the maximum distance) because of the different terms used. The cosine distance between the two sentence-BERT representations, by contrast, is 0.209, capturing the semantic relationship between the terms. Similarly, the two businesses “importing all household goods” and “cross border trading” share no terms in common, giving them a cosine distance value of 1.0 in the standard approach, despite the fact that both enterprises are related to import-export functions. The sentence-BERT measure of cosine distance better captures that relatedness, giving the two businesses a pairwise distance of 0.513.

An additional advantage of the multilingual sentence-BERT models is that they are able to compute semantic similarity between phrases in different languages (Reimers and Gurevych, 2020a), as demonstrated in the last example of Table 5. This is useful when interviews are
Table 5: Comparison of Standard Cosine Distance with Sentence-BERT Embeddings

<table>
<thead>
<tr>
<th>Business 1</th>
<th>Business 2</th>
<th>Pairwise Cosine Distance (Bag-of-Words)</th>
<th>Pairwise Cosine Distance (SBERT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hair salon</td>
<td>Barber</td>
<td>1.000</td>
<td>0.209</td>
</tr>
<tr>
<td>Sell chickens</td>
<td>Poultry</td>
<td>1.000</td>
<td>0.174</td>
</tr>
<tr>
<td>Sell chickens</td>
<td>Sell chicks</td>
<td>0.500</td>
<td>0.179</td>
</tr>
<tr>
<td>Dressmaking</td>
<td>Tailor</td>
<td>1.000</td>
<td>0.269</td>
</tr>
<tr>
<td>Importing all household goods</td>
<td>Cross border trading</td>
<td>1.000</td>
<td>0.513</td>
</tr>
<tr>
<td>Buying and selling of perfumes</td>
<td>Selling cosmetics</td>
<td>0.684</td>
<td>0.244</td>
</tr>
<tr>
<td>Repairing phones and selling them</td>
<td>Electronics market</td>
<td>1.000</td>
<td>0.397</td>
</tr>
<tr>
<td>Growing vegetables, tomatoes, maize, and poultry</td>
<td>Grow green vegetables, tomatoes, beans, and do poultry as well</td>
<td>0.484</td>
<td>0.098</td>
</tr>
<tr>
<td>Nabebenta ng mga sari saring pagkain*</td>
<td>Selling food</td>
<td>1.000</td>
<td>0.408</td>
</tr>
</tbody>
</table>

*“Selling a variety of foods” in Filipino, an example from Karlan and Zinman (2011).*

conducted in a mix of languages (for example, alternating between English and a local language) and different survey-takers record responses in different languages. Using an example from one of the meta-analysis studies described in the following section (Karlan and Zinman, 2011), in which the business descriptions are recorded in a mix of English and Filipino, the table demonstrates how the sentence-BERT model is able to capture the relatedness between a Filipino phrase meaning “Selling a variety of foods” and the English phrase “Selling food”, whereas the standard word vector approach would naturally not recognize the similarity because of the lack of common terms used.

A full analysis using differentiation measures based on the bag-of-words vector representations can be found in the appendix.

4 Results

Table 6 displays summary statistics within datasets and correlation coefficients for the combined sample. There are considerable differences between the populations and contexts of
the component studies. The businesses in the Bangladesh and Mexico studies were virtually all male-owned, for example, while the businesses in the Kenya and Mongolia studies were almost entirely female-owned. Mean years of schooling ranged from a scant 5.1 years, in the Uganda study, to 13.7 years, in the Philippines study. Business size ranged from tiny – the microenterprises in the Uganda study reported the equivalent of 25 USD in revenues per month – to more robust small businesses: the businesses in the Bangladesh sample employed 2.75 people on average and reported over 11000 USD in revenues per month. Most business owners were in their late 30s or early 40s on average at the time of data collection, with the exception of the Zimbabwe sample, which specifically targeted younger entrepreneurs.

Table 6: Summary Statistics and Correlations

<table>
<thead>
<tr>
<th>Summary Statistics</th>
<th>Differentiation Score</th>
<th>Female</th>
<th>Age</th>
<th>Years of Schooling</th>
<th>Total Employment</th>
<th>Monthly Revenue (USD)</th>
<th>Monthly Profit (USD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined Dataset</td>
<td>-0.004 (1.00)</td>
<td>0.652</td>
<td>37.9 (10.2)</td>
<td>9.2 (4.0)</td>
<td>0.94 (1.92)</td>
<td>2438 (16753)</td>
<td>243 (1110)</td>
</tr>
<tr>
<td>Mongolia 2008</td>
<td>0.00 (1.00)</td>
<td>0.981</td>
<td>41.1 (9.2)</td>
<td>9.1 (2.7)</td>
<td>-</td>
<td>2367 (3897)</td>
<td>-</td>
</tr>
<tr>
<td>Philippines 2008</td>
<td>0.00 (1.00)</td>
<td>0.838</td>
<td>41.9 (8.9)</td>
<td>13.7 (2.1)</td>
<td>0.71 (1.54)</td>
<td>1219 (1834)</td>
<td>363 (718)</td>
</tr>
<tr>
<td>Uganda 2007</td>
<td>0.00 (1.00)</td>
<td>0.522</td>
<td>34.2 (10.8)</td>
<td>5.1 (3.6)</td>
<td>-</td>
<td>25 (68)</td>
<td>11 (32)</td>
</tr>
<tr>
<td>Togo 2013</td>
<td>0.00 (1.00)</td>
<td>0.526</td>
<td>41.2 (9.7)</td>
<td>6.8 (3.5)</td>
<td>-</td>
<td>1516 (5805)</td>
<td>198 (417)</td>
</tr>
<tr>
<td>Bangladesh 2010</td>
<td>0.00 (1.00)</td>
<td>0.018</td>
<td>41.5 (10.9)</td>
<td>9.9 (4.4)</td>
<td>2.75 (2.98)</td>
<td>11315 (39719)</td>
<td>775 (2466)</td>
</tr>
<tr>
<td>Mexico 2005</td>
<td>0.00 (1.00)</td>
<td>0.000</td>
<td>37.0 (9.4)</td>
<td>6.5 (4.1)</td>
<td>1.24 (0.64)</td>
<td>667 (419)</td>
<td>-</td>
</tr>
<tr>
<td>Kenya 2013</td>
<td>0.00 (1.00)</td>
<td>0.997</td>
<td>35.7 (9.0)</td>
<td>9.0 (2.9)</td>
<td>0.27 (0.62)</td>
<td>210 (246)</td>
<td>44 (67)</td>
</tr>
<tr>
<td>Zimbabwe 2017</td>
<td>0.00 (1.00)</td>
<td>0.549</td>
<td>28.3 (4.4)</td>
<td>12.1 (2.0)</td>
<td>0.34 (1.03)</td>
<td>500 (819)</td>
<td>187 (255)</td>
</tr>
</tbody>
</table>

Correlations: Combined Dataset

<table>
<thead>
<tr>
<th>Differentiation Score</th>
<th>Female</th>
<th>Age</th>
<th>Years of Schooling</th>
<th>Employment</th>
<th>Revenue</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differentiation Score</td>
<td>-0.063</td>
<td>-0.061</td>
<td>-0.037</td>
<td>-0.078</td>
<td>0.018</td>
<td>0.031</td>
</tr>
</tbody>
</table>

Top panel displays means with standard deviations in parentheses. Bottom panel displays Pearson correlation coefficients. Within-dataset correlations can be found in the Appendix.

5 The Bangladesh sample included even larger businesses that were dropped from this analysis; results with those larger businesses can be found in the appendix in an analysis of what constitutes “micro” enterprise.
4.1 Baseline Relationship Between Differentiation and Microenterprise Performance

Within each of the meta-analysis samples, the Differentiation Score was calculated using the text business descriptions as described above. The differentiation measures were calculated only within sample; that is, only with reference to their peer businesses in the same study. I then examined the relationships between the text measures of differentiation and the performance measures, when available. In each case, I show the results in a forest plot, in a pseudo-meta-analysis in which the relationship was calculated within each dataset with an OLS regression of the form

\[ Y_i = \beta_0 + \beta_1 \text{DifferentiationScore}_i + \beta X_i + \epsilon_i \]

in which the dependent variable \( Y \) in each model is the mean-adjusted performance measure (e.g. revenues), and the covariate vector \( X \) controls for the age, gender, and years of schooling of the primary owner, as well as the sector of the microenterprise (retail, services, manufacturing, agriculture, or other). As the performance variables for the Uganda, Philippines, and Zimbabwe datasets are post-treatment, these additionally control for experimental treatment status; all others are baseline values (pre-treatment).

In each case, I also display a binned scatterplot\(^6\) of the relationship between the performance measure and the differentiation score using the combined dataset, in which a fixed effect for each individual study is added to the previously mentioned control variables.

Figure 1 displays the relationship between the differentiation score and monthly firm revenues. As shown in the forest plot, the relationship is positive in six of the eight subsamples and reaches statistical significance in three of eight. The overall random effects analysis

\(^6\)Created using the \textit{binscatter} package in Stata, with the default bins of 20 quantiles unless otherwise specified.
estimates that a one-standard-deviation increase in the differentiation score is associated with an 11 percent increase in revenue relative to the mean (95 percent confidence interval: 0.03 – 0.20). The binned scatterplot suggests that this relationship may not be purely linear, however; it appears to peak around approximately one standard deviation above the mean before declining slightly.

Figure 2 displays the relationship between the differentiation score and monthly profits. One again, the forest plot shows that a positive relationship can be found in a majority of the subsamples. The overall random effects analysis estimates that a one-standard-deviation increase in the differentiation score is associated with an eight percent increase in profits relative to the mean (95 percent confidence interval: 0.04 – 0.12). Again, the binned scatterplot suggests that the relationship peaks at approximately one standard deviation above the mean level of differentiation before declining, raising the question of whether there may be a point of optimal distinctiveness for microenterprises (Deephouse, 1999; Zhao et al., 2017). We return to this question later on in the results and the subsequent discussion.

Finally, Figure 3 displays the relationship between the differentiation score and mean-
Overall Relationship Between Differentiation and Firm Profit

Left panel displays estimated coefficients from OLS regressions of mean-adjusted monthly profits on the Differentiation Score within each dataset. The random effects analysis estimates that a one-standard-deviation increase in the differentiation score is associated with an eight percent increase in profit relative to the mean. Each model controls for proprietor age, gender, years of education, and enterprise sector (retail, services, manufacturing, agriculture, or other). Right panel displays a binned scatterplot of the relationship between monthly profits and the Differentiation Score within the combined dataset, controlling for the same covariates as in the left-hand panel and with an additional fixed effect for each individual study.

Adjusted total employment. I display the results both with and without controlling for sector as employment values tend to be highly correlated with sector (service microenterprises have much higher employment levels than retail microenterprises, for example). Therefore, while the differentiation score appears to be highly correlated with employment levels without sector controls – a one-standard-deviation increase in the differentiation score is associated with a 20 percent increase in employment relative to the mean in this estimate (95 percent confidence interval: 0.03 – 0.38) – the estimated relationship shrinks substantially once adjusted for sector effects. Controlling for sector, a one-standard-deviation increase in the differentiation score is associated with a 10 percent increase in employment (95 percent confidence interval: −0.03 – 0.24). Again, both binned scatterplots display a pattern of non-linearity with a peak approximately one standard deviation above the mean.
(a) Without Controlling for Sector

(b) Controlling for Sector

**Figure 3: Overall Relationship Between Differentiation and Employment**

Left panels display estimated coefficients from OLS regressions of mean-adjusted total employment on the *Differentiation Score* within each dataset. Each model controls for proprietor age, gender, and years of education. The bottom panel additionally controls for enterprise sector (retail, services, manufacturing, agriculture, or other). The random effects analysis estimates that a one-standard-deviation increase in the differentiation score is associated with a 20 percent increase in employment relative to the mean; however, this relationship shrinks to a 10 percent increase once adjusted for sector effects. Right panels display a binned scatterplot of the relationship between employment and the *Differentiation Score* within the combined dataset, controlling for the same covariates as in the left-hand panel and with an additional fixed effect for each individual study.
4.2 Differentiation Within and Across Sectors

As demonstrated in the previous section, business sector is an important variable in understanding the relationship between differentiation and performance. The majority of the microenterprises in the combined sample – 64 percent – are retail businesses. The second-largest sector is services, comprising 22 percent of the sample. Manufacturing and agriculture businesses make up seven and five percent of the combined sample, respectively, and the remaining two percent are categorized as other.

Figure 4 displays summary statistics of the differentiation score within each sector. Retail businesses have the lowest differentiation scores on average, while services businesses and those categorized as other have the highest scores.

**Figure 4: Differentiation Score Across Sectors**

![Boxplot of Differentiation Score Across Sectors](image)

Plot displays summary statistics of the standardized *Differentiation Score* within each of the five primary sectors. The lower and upper hinges of each box correspond to the first and third quartiles (the 25th and 75th percentiles), while the dark center line represents the median.

Figure 5 displays the estimated relationship between differentiation and performance measures within each sector (mean-adjusted monthly revenue and profit in the left and right panels, respectively). Displayed for each sector is the coefficient $\beta_1$ from an OLS regression of the specification
estimated within each sector, in which $\alpha_j$ is a fixed effect for each individual study $j$ in the combined dataset and the covariate vector $X$ controls for the age, gender, and years of schooling of the primary owner. The models are more precisely estimated in the two most common sectors, retail and services, and display an interesting pattern. Within retail, the positive relationship between differentiation and revenues ($\beta = 0.095, p = 0.0004$) and differentiation and profits ($\beta = 0.070, p = 0.027$) resembles the estimates for the overall sample. Within the services businesses, however, there appears to be no relationship between differentiation and performance ($\beta = -0.011, p = 0.676$ for revenue and $\beta = 0.018, p = 0.639$ for profit). Both the manufacturing and agriculture sectors demonstrate evidence of a positive association between differentiation and performance, though the estimates are less precise due to the smaller sample sizes in those sectors. Interestingly, there appears to be a negative relationship between differentiation and performance within businesses classified as “other”, though as this category is very small and functions as a catch-all for all firms that do not fall into a previously mentioned sector, it is hard to interpret this finding in a meaningful way.
Figure 5: Relationship Between Differentiation and Performance, Within Sector

Plot displays estimated coefficients from OLS regressions of mean-adjusted revenues (left panel) and profits (right panel) on the Standard Differentiation Score, within each sector. Each model controls for proprietor age, gender, and years of education and includes a fixed effect for each individual study. Bars display 95-percent confidence intervals.

4.2.1 Within Retail

Given that the canonical microenterprise is a small retail business, and that the bulk of the microenterprises in the combined sample are retail firms, I next replicate the primary analysis within the retail sector. Figure 6 displays both forest plots and binned scatterplots of the association between the differentiation score and revenue, profit, and employment for retail firms only. The random effects analysis estimates that a one-standard-deviation increase in the differentiation score is associated with a 12 percent increase in revenue (95 percent confidence interval: 0.04 – 0.19), a 13 percent increase in profit (95 percent confidence interval: −0.03 – 0.29), and a 14 percent increase in employment (95 percent confidence interval: −0.03 – 0.33) relative to the mean, respectively.
Figure 6: Relationship Between Differentiation and Performance, Retail Sector

Left panels display estimated coefficients from OLS regressions of mean-adjusted revenues (top panel), profits (second panel), and employment (third panel), respectively, on the Differentiation Score within each dataset for retail businesses only (N = 6377). Each model controls for proprietor age, gender, and years of education. The random effects analysis estimates that a one-standard-deviation increase in the differentiation score is associated with a 12 percent increase in revenue, a 13 percent increase in profit, and a 15 percent increase in employment relative to the mean, respectively. Right panels display a binned scatterplot of the relationship between each performance measure and the Differentiation Score within the combined dataset, controlling for the same covariates as in the left-hand panel and with an additional fixed effect for each individual study.
4.3 Who is Able to Differentiate? Men, the Young, and Those With More Schooling

The next section of results will examine in detail how the individual characteristics of the primary proprietor of the microenterprise affect propensity to differentiate. Figure 7 displays binned scatterplots of the relationship between the differentiation score and proprietor age or years of schooling, respectively.

The binned scatterplot suggests that propensity to differentiate falls linearly with the age of the proprietor. A linear regression estimates that an increase in age of ten years is associated with a decline in the Differentiation Score of 0.068 standard deviations ($p < 0.001$). By contrast, the propensity to differentiate rises with the proprietor’s years of schooling. The linear regression estimates that each year of schooling is associated with an increase in the Differentiation Score of 0.029 standard deviations ($p < 0.001$), though the binned scatterplot displays evidence of a discontinuity around 12 years of schooling (that is, there is a jump in the differentiation scores for those with any post-secondary education).

Figure 8 displays density plots of the distribution of the differentiation scores within female-owned businesses (dark solid line) and male-owned businesses (lighter dotted line). Both genders have a skewed distribution, with a peak of less-differentiated businesses below the mean and a longer right tail of more differentiated businesses. This skew is more pronounced for female microenterprise owners, however, with a higher density of businesses concentrated below the mean level of differentiation. This distribution pattern results in female-owned businesses having on average 0.336 standard deviations lower Differentiation Scores than male-owned businesses ($p < 0.001$, controlling for individual study fixed effects).

The relationship between differentiation and performance within female-owned versus male-owned businesses sheds additional light on this finding. Figure 9 displays binned scatterplots of the relationship between the Differentiation Score and performance measures for female-
Figure 7: Propensity to Differentiate by Age and Years of Schooling

Figure displays binned scatterplots of the relationship between the Differentiation Score and proprietor age (left panel) and proprietor years of schooling (right panel) in the combined dataset, controlling for the fixed effect of each individual study. The linear best-fit lines estimate that an increase in age of ten years is associated with a decline in the Differentiation Score of 0.068 standard deviations ($p < 0.001$), and that each additional year of schooling is associated with an increase in the Differentiation Score of 0.029 standard deviations ($p < 0.001$).

Figure 8: Distribution of Differentiation Score by Gender

Plot displays kernel density estimates of the distribution of the Differentiation Score within female and male proprietors across all datasets. Controlling for individual study fixed effects, female-owned businesses have on average 0.336 standard deviations lower Differentiation Scores than male-owned businesses ($p < 0.001$).
owned businesses (left panels) and male-owned businesses (right panels). As in prior analyses, estimates control for age, education, sector, and individual study fixed effects.

The differences are pronounced. The estimated relationship between the *Differentiation Score* and revenue for women entrepreneurs ($\beta = 0.026, p = 0.140$) is approximately 8.7 times smaller in magnitude than the corresponding relationship for men ($\beta = 0.230, p = 0.0002$). A similar pattern can be found in the relationship between differentiation and profit, with the estimated relationship for women ($\beta = 0.031, p = 0.245$) being 6.0 times smaller in magnitude than its counterpart for men ($\beta = 0.185, p = 0.002$). The results suggest that female proprietors are not only less likely to have highly differentiated microenterprises, but that differentiation is less closely linked to performance for women than it is for men. The binned scatterplots also suggest that the appearance of a point of optimal differentiation discussed in earlier Figures 1-3 may be primarily driven by the female-owned businesses: for male microenterprise owners, the most highly differentiated businesses (at or near two standard deviations above the mean) have some of the highest levels of revenues and profits, whereas the same performance benefits cannot be observed among their female-owned counterparts.
Figure 9: Relationship Between Differentiation and Performance, By Gender

Plots display binned scatterplots of the relationship between mean-adjusted revenues (top panels) and profits (bottom panels) with the Differentiation Score. Relationships within female-owned businesses ($N = 6367$) are displayed on the left, while relationships within male-owned businesses ($N = 3387$) are displayed on the right. The relationship between revenue and differentiation is approximately 8.7 times stronger for men than for women, while the relationship between profit and differentiation is approximately 6.0 times stronger. Estimates control for proprietor age, years of education, and enterprise sector (retail, services, manufacturing, agriculture, or other), as well as including a fixed effect for each individual study.
4.4 Impact of Policy Interventions on Differentiation

The final section of results examines whether or not any of the randomized interventions conducted in the individual studies had causal impacts on the Differentiation Score. The majority of the component datasets involved some type of randomized experiment. The interventions fell into two broad categories: four studies included some type of business training program, while three studies provided or encouraged access to a microloan product. In each case, Tables 7 and 8 display coefficient plots with the estimates from an OLS regression of the Differentiation Score on assignment to the relevant treatment condition(s).

Each of the training programs differed slightly in their specifics, though all were relatively limited in scope: three of the four comprised five days of training, while the fourth included 36 hours of training spread over four weeks. The program in Uganda (Blattman et al., 2016) included an additional cash grant worth 150 USD, as well as ongoing supervision for program recipients. The intervention in Togo (Campos et al., 2017) included two variations on training, one more traditional version of business training and another version that encouraged the formation of personal initiative and other goal-setting behaviors. The program in Kenya (McKenzie and Puerto, 2021a) was specifically tailored for female entrepreneurs and also tested for market-level training spillover effects. The intervention in Zimbabwe included a referral to a microloan organization in addition to the business training (Carlson and Hager, 2021).

The estimated treatment effects of each of the training programs on the Differentiation Score are, without exception, remarkably close to zero. The time range in which outcomes were collected ranges from nine months to two years post-treatment. In the case of the Togo study, outcomes were measured four times post-treatment over the course of two years. The results suggest that the types of standard, time-limited business training interventions aimed at microenterprise owners do not increase their propensity to differentiate.
### Table 7: Do Training Programs Increase Differentiation?

<table>
<thead>
<tr>
<th>Study</th>
<th>Description of Intervention</th>
<th>Estimated Treatment Effect on Differentiation Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uganda 2007</td>
<td>$150 cash grant, five days of business skills training (business planning, sales, marketing, record-keeping, and budgeting), and ongoing supervision. Outcomes collected 16 months following the cash grant.</td>
<td>Training + Grant <img src="image" alt="Coefficient plot" /></td>
</tr>
<tr>
<td>Togo 2013</td>
<td>36 hours of training over four weeks, either traditional business training (&quot;Managerial Training&quot;) – finance, marketing, human resource management, and formalization – or personal initiative business training (&quot;Entrepreneurial Training&quot;) – self-starting behavior, goal-setting, planning, and overcoming obstacles. Outcomes collected at four endpoints over the two years following treatment.</td>
<td><img src="image" alt="Coefficient plot" /></td>
</tr>
<tr>
<td>Kenya 2013</td>
<td>Five days of business skills training designed specifically for female business owners (topics on gender and entrepreneurship, business planning, marketing, bookkeeping, and soft skills). “Market-Level Training” indicates that overall market was selected for training to measure any prospective spillovers, while “Individual-Level Training” indicates that the individual microenterprise owner was assigned to training. Outcomes collected one year after treatment.</td>
<td><img src="image" alt="Coefficient plot" /></td>
</tr>
<tr>
<td>Zimbabwe 2017</td>
<td>Five days of business skills training (business planning, sales, marketing, record-keeping, and ethics). Outcomes collected 9-10 months after treatment.</td>
<td><img src="image" alt="Coefficient plot" /></td>
</tr>
</tbody>
</table>

Coefficient plots display estimated average treatment effect of assignment to respective intervention on the Differentiation Score derived from OLS regressions without adjusting for covariates. Bars display 95 percent confidence intervals. Togo plot displays results from four separate regression models (one for each of four follow-up surveys).
Table 8 displays a similar analysis that examines whether increasing access to microloan products has any treatment effects on differentiation. Three studies included an intervention that served in some way to increase access to microfinance. Attanasio et al. (2015) compared the impact of increasing access to joint-liability loans versus individual-liability loans for rural Mongolian women. Karlan and Valdivia (2011) introduced a credit-scoring product to randomly assign individual-liability loans in the Philippines. Finally, the training program conducted in Carlson and Hager (2021) included a referral to a microloan organization to encourage use of individual-liability loans, though only a minority of participants elected to take up the loan.

The results are suggestive. In the only case in which access to a joint-liability loan was randomly offered, the treatment appears to have had a null effect on differentiation. The individual-liability loan, however, appears to have had a negative effect on the differentiation score, consistent with the effect observed in Zimbabwe (though, as the Zimbabwe study displays the effect of actually receiving a microloan, rather than random assignment, it should be interpreted with some caution). The estimated treatment effect of the individual-liability loans provided in the Philippines is not significantly different from zero, though the loans in that study were much smaller in magnitude than in the other two studies. In both cases in which data on the size of the loan requested was available (a value endogenous to the loan requestor), the value of the loan was positively associated with the Differentiation Score post-treatment.
Table 8: Does Loan Access Increase Differentiation?

<table>
<thead>
<tr>
<th>Study</th>
<th>Description of Intervention</th>
<th>Estimated Treatment Effect on Differentiation Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mongolia 2008</td>
<td>Access to joint-liability (average size $279) or individuality-liability (average size $411), monthly interest rate 1.5 to 2 percent, monthly repayment. Outcomes collected 19 months after treatment.</td>
<td>![Coefficient plot for Mongolia 2008]</td>
</tr>
<tr>
<td>Philippines 2008</td>
<td>Access to individual-liability loans (average size $225), monthly interest rate 2.5 percent, weekly repayment. Outcomes collected 11 to 22 months after treatment.</td>
<td>![Coefficient plot for Philippines 2008]</td>
</tr>
<tr>
<td>Zimbabwe 2017</td>
<td>Referral to individual-liability loans (only 19 percent takeup, average size $495), monthly interest rate 1.5 percent, monthly repayment. Outcomes collected 9-10 months after treatment.</td>
<td>![Coefficient plot for Zimbabwe 2017]</td>
</tr>
</tbody>
</table>

Coefficient plots display estimated average treatment effect of respective intervention on the Differentiation Score derived from OLS regressions without adjusting for covariates. Mongolia and Philippines studies display the effect of assignment to treatment condition, while Zimbabwe study displays effect of loan take-up. Bars display 95 percent confidence intervals.

5 Discussion

The results across the individual component datasets suggest that the relationship between differentiation and microenterprise performance is fairly robust, though the nature of the association does vary quite a bit along the lines of things like firm sector and proprietor characteristics. Particularly within the dominant sector, retail – close to two-thirds of the businesses in the combined dataset – the evidence demonstrates that more differentiated businesses are consistently not just larger, but more profitable. The fact that the association between differentiation and performance is absent in the second-largest sector, service businesses, is curious. It may be that the service sector is less saturated with microenterprises than the others, or that small service businesses – which tend to generate more employment than the other sectors – differ in other systematic ways.

The differences in microentrepreneurs’ individual propensity to differentiate raise many interesting questions. The negative relationship between differentiation and age, in particular,
suggests there may be cognitive factors at play that limit the potential to conceive of or execute a differentiated offering. The findings are in line with research on creativity and age, suggesting that youth is the best time for paradigm-breaking ideas (Azoulay et al., 2019; Dietrich and Srinivasan, 2007). The association between differentiation and years of schooling also suggests that a lack of human capital may limit the propensity to differentiate, although this relationship is likely to be confounded by factors such as socioeconomic status and family income.

The gender results are particularly striking. The findings suggest not only that women are less likely to differentiate, but that they seem to see less benefit to differentiating when they do so. Several factors may be acting upon these results: first, other well-studied constraints to women’s enterprise growth – e.g., capture of income by family members and male partners, cultural norms, childcare responsibilities, lack of access to capital (Jayachandran, 2020) – may be the more binding constraint with respect to firm growth, relative to strategic positioning. Another (not mutually exclusive) possibility is that women proprietors may face discrimination from customers or suppliers if they strike out with a business idea that is “outside the box” or does not fit with expected norms of feminine behavior (Heilman, 1983, 2012). This is consistent with prior work showing women business owners are more rewarded when their offering is consistent with female stereotypes (Lee and Huang, 2018), and may reduce both the propensity to differentiate in the first instance, as well as the positive relationship between differentiation and performance observed in the male-owned microenterprises.

Finally, while the evaluation of the randomized interventions may seem to have produced disappointing results, the findings are informative. It seems fairly clear that the types of existing business training programs aimed at microenterprises do little to increase differentiation (though it’s worth noting that all the programs studied targeted existing entrepreneurs and were quite limited in duration). Additionally, the findings suggest that individual-liability
microloan access may decrease the propensity to differentiate. This is consistent with experimental findings that individual-liability loans – which place more repayment responsibility on the shoulders of the individual borrower – are less likely to encourage risk-taking than joint-liability loans (Giné et al., 2010). At the same time, the relationship between requested loan size and differentiation suggests that highly differentiated businesses may have greater capital needs (although this result also consistent with the finding, established earlier, that more differentiated businesses tend to be larger and more profitable).

The findings in this analysis establish only an association, of course, not a definitive causal relationship running from differentiation to performance. Nevertheless, the results warrant an attempt at finding an intervention that, in contrast to the ones studied above, may positively shift microentrepreneurs’ propensity to differentiate. The ideal intervention to pilot would be one that applies maximal pressure to increase differentiation outcomes – perhaps by bundling multiple aspects of treatment – allowing researchers to observe if performance outcomes do indeed follow causally through that channel. If a successful intervention were to be found, subsequent pilots could then more precisely identify the relevant mechanisms of treatment.

Given the findings discussed above, I suggest several considerations in designing such an intervention: first, it would likely be most effective to target new entrepreneurs, as opposed to those with existing businesses – or, perhaps, recruit entrepreneurs who explicitly want to move into a new business area. Second, it may be most impactful to pilot the intervention first on male microenterprise owners, given that the relationship between differentiation and performance was observed to be much stronger for men than for women. It is likely that other constraints to women’s business growth may need to be loosened first. Similarly, I would suggest piloting the intervention solely on retail businesses, given that the relationship between differentiation and performance appears most robust in this sector. Third, any type of training intervention should include a stronger emphasis on strategy and positioning.
concepts, helping entrepreneurs to identify in particular where they might source their differentiated offering – supplier relationships are likely to matter, for example. Longer programs may be also more effective. Finally, while financial resources are likely to be an important factor, loans with harsh repayment terms may decrease risk-taking and limit propensity to differentiate. Cash grants or loans with flexible terms may be more effective.

6 Conclusion

This article examines the role of differentiation in small informal businesses in developing economies. To do so, it introduces a text-based measure of differentiation, capturing the extent to which a given business is distinct from its peers in the sample. The method uses sentence embedding vectors from the sentence-BERT embeddings model, offering a way to capture business model similarity that is less likely to be sensitive to inexact matches in language use than bag-of-words approaches and allows for computation across multiple languages.

In an analysis comprising over 10,000 total microenterprises across eight countries, the text-based measure of differentiation is positively associated with business performance. The analysis suggests that a standard deviation increase in differentiation from peer businesses is associated with an increase of 11 percent in revenue and eight percent in profits relative to the mean, a relationship that is present in retail, agriculture, and manufacturing businesses, though not in service enterprises. Individual propensity to engage in differentiation varies with the characteristics of the proprietor: younger, male, and more highly educated microenterprise owners are more likely to have a highly differentiated businesses, relative to older, female, and less schooled individuals, respectively. The relationship between differentiation and performance is also substantially stronger for men than for women. Examining the randomized interventions implemented in the individual component studies, the results suggest that business training programs do not impact microenterprises’ level of differentia-
tion, and that increasing access to individual-liability loan products may decrease propensity to differentiate.

By introducing a computationally simple text-based measure of distance, applicable to various types of text data, the article provides a methodological tool that can capture differentiation within a sample of businesses. At a minimum, this measure represents a simple way of controlling for a microenterprise’s degree of differentiation, for researchers who are interested in studying other constructs. The analysis also demonstrates the usefulness of sentence embedding models in providing robustness to standard text-based distance measures, showing through tractable examples the ways in which embeddings can guard against some of the pitfalls associated with measuring the similarity of texts. Sentence embeddings should prove useful for a number of other applications within strategy and management research. The method is particularly suitable for the length of text sources such as newspaper headlines, tweets, titles (e.g., the section headings in annual letters to shareholders, the titles in crowdfunding projects), or company mottos and slogans, for example. The methods used here may need to be modified slightly to account for longer texts.

The empirical exercise in this article sheds light on a relatively understudied population of businesses within the strategy literature, building upon work on enterprise under conditions of poverty, scarce resources, and institutional voids (e.g., Mair and Marti, 2009; Bruton et al., 2013; Alvarez and Barney, 2014; Sutter et al., 2019). There is scope for a good deal of future work that builds upon the findings in this study. A natural extension would be through studies that explicitly target text-based measures of differentiation as an experimental outcome, testing interventions along the lines discussed in the previous section. Other extensions might investigate some of the questions raised by the findings on differentiation across sector and individual characteristics: for example, how age impacts strategic decisions made by small business owners, whether female enterprise owners are discriminated against for differentiating, and the nature of strategic positioning within micro service en-
terprises. While the types of businesses studied in this analysis may be tiny, their combined economic contribution is enormous, employing nearly two billion people worldwide (International Labour Organization, 2019): a better understanding of the internal dynamics of micro enterprises may provide insight on how to improve the livelihoods of many.
References


Balloon Ventures (2018). Decent Work in Uganda’s Informal Economy.


Sutter, C., Bruton, G. D., and Chen, J. (2019). Entrepreneurship as a Solution to Ex-


Appendix

Sample Python Code

**Differentiation Measure**

```python
# transformer model
model = SentenceTransformer('paraphrase-multilingual-mpnet-base-v2')
# note: full list of models can be found at https://www.sbert.net/docs/pretrained_models.html

# sentences are encoded by calling model.encode(): this converts each sentence to an embedding vector of length 768
sentences = list(all_data['business_activity'])
embeddings = model.encode(sentences)

# cosine scores: this loops through and computes all pairwise cosine distance scores, then takes the mean
cosim_scores = [0] * len(sentences)
for i in range(0, len(sentences)):
    cosim_score = [0] * len(sentences)
    for j in range(0, len(sentences)):
        cosim_score[j] = 1 - distance.cosine(embeddings[i], embeddings[j])
    cosim_scores[i] = np.mean(cosim_score)

# add final scores to data frame
all_data['dissertation_score'] = cosim_scores
```

Code written in Python 3.8.2. Full code and sample data can be found at [link redacted for review].

**Figure A1:** Distribution of Differentiation Score

Plot displays kernel density estimates of the distribution of the Differentiation Score across all datasets.
(a) Without Controlling for Revenue and Employment  
(b) Controlling for Revenue and Employment

Figure A2: Overall Relationship Between Differentiation and Profitability, Accounting for Size

Left panel displays a binned scatterplot of the relationship between mean-adjusted monthly profits and the Differentiation Score within the combined dataset, identical to the plot in Figure 2 (controlling for proprietor age, gender, years of education, enterprise sector, and with a fixed effect for each individual study). Right panel replicates this exercise and additionally controls for mean-adjusted monthly revenues and total employment. The magnitude of the estimated relationship shrinks by 25 percent but remains statistically significant ($p = 0.002$ versus $p = 0.009$).
## Table A1: Individual Dataset Correlations

<table>
<thead>
<tr>
<th>Country</th>
<th>Differentiation Score</th>
<th>Female</th>
<th>Age</th>
<th>Years of Schooling</th>
<th>Employment</th>
<th>Revenue</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mongolia 2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differentiation Score</td>
<td>-0.091</td>
<td>0.079</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.245</td>
<td>0.026</td>
<td>0.125</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Schooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>0.132</td>
<td>-0.099</td>
<td>-0.016</td>
<td>0.071</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Philippines 2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differentiation Score</td>
<td>-0.118</td>
<td>0.014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.121</td>
<td>0.112</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Schooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>0.175</td>
<td>-0.066</td>
<td>-0.048</td>
<td>0.127</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue</td>
<td>0.011</td>
<td>-0.012</td>
<td>-0.035</td>
<td>0.073</td>
<td>0.321</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td>0.053</td>
<td>-0.037</td>
<td>-0.021</td>
<td>0.087</td>
<td>0.271</td>
<td>0.623</td>
<td></td>
</tr>
<tr>
<td>Uganda 2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differentiation Score</td>
<td>-0.097</td>
<td>0.014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.064</td>
<td>0.167</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Schooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>0.173</td>
<td>-0.355</td>
<td>-0.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue</td>
<td>0.148</td>
<td>-0.115</td>
<td>0.042</td>
<td>0.145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td>0.019</td>
<td>-0.074</td>
<td>0.085</td>
<td>0.078</td>
<td>0.707</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Togo 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differentiation Score</td>
<td>-0.241</td>
<td>0.006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.066</td>
<td>0.167</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Schooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue</td>
<td>0.016</td>
<td>-0.042</td>
<td>-0.042</td>
<td>0.023</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td>0.018</td>
<td>-0.087</td>
<td>-0.020</td>
<td>-0.015</td>
<td>0.526</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangladesh 2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differentiation Score</td>
<td>0.026</td>
<td>-0.046</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.024</td>
<td>-0.024</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Schooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>0.157</td>
<td>-0.014</td>
<td>0.136</td>
<td>0.132</td>
<td>0.186</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue</td>
<td>-0.040</td>
<td>0.070</td>
<td>0.139</td>
<td>0.186</td>
<td>0.200</td>
<td>0.498</td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td>-0.001</td>
<td>-0.023</td>
<td>0.083</td>
<td>0.108</td>
<td>0.200</td>
<td>0.498</td>
<td></td>
</tr>
<tr>
<td>Mexico 2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differentiation Score</td>
<td>-0.106</td>
<td>0.038</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.179</td>
<td>-0.179</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Schooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>0.078</td>
<td>-0.040</td>
<td>0.129</td>
<td>0.048</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kenya 2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differentiation Score</td>
<td>0.022</td>
<td>0.016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.000</td>
<td>-0.015</td>
<td>-0.202</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Schooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>0.147</td>
<td>-0.005</td>
<td>-0.013</td>
<td>0.092</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue</td>
<td>0.013</td>
<td>0.020</td>
<td>-0.099</td>
<td>0.13</td>
<td>0.074</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td>0.044</td>
<td>0.008</td>
<td>0.001</td>
<td>0.088</td>
<td>0.119</td>
<td>0.175</td>
<td></td>
</tr>
<tr>
<td>Zimbabwe 2017</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Differentiation Score</td>
<td>-0.176</td>
<td>0.011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.061</td>
<td>0.054</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of Schooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>0.112</td>
<td>-0.019</td>
<td>0.012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revenue</td>
<td>0.123</td>
<td>-0.107</td>
<td>0.087</td>
<td>0.201</td>
<td>0.513</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Profit</td>
<td>0.091</td>
<td>-0.111</td>
<td>0.062</td>
<td>0.197</td>
<td>0.361</td>
<td>0.809</td>
<td></td>
</tr>
</tbody>
</table>

Panels display Pearson correlation coefficients.
Results with Bag-of-Words Measure

Table A2: Correlations Between SBERT Differentiation Score and Bag-of-Words Differentiation Score

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Correlation Coefficient</th>
<th>Examples of False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined Dataset</td>
<td>0.660</td>
<td>“Crafts carving furniture making”</td>
</tr>
<tr>
<td>Mongolia 2008</td>
<td>0.955</td>
<td>“General merchandise”; “Selling variety”; Filipino texts and texts with typos, e.g. “Grocerysoftdrinks etc”, “Gulay prutas” (vegetables fruits)</td>
</tr>
<tr>
<td>Philippines 2008</td>
<td>0.606</td>
<td>Typos and untranslatable Acholi words, e.g. “Cotton charcoal”, “Chacoal”, “Kweteyenum ot”</td>
</tr>
<tr>
<td>Uganda 2007</td>
<td>0.584</td>
<td>“Menuiser” (carpenter); “Bâtiments et fourniture” (buildings and supplies)</td>
</tr>
<tr>
<td>Togo 2013</td>
<td>0.501</td>
<td>“Other wholesale”; “Other manufacturing”</td>
</tr>
<tr>
<td>Bangladesh 2010</td>
<td>0.619</td>
<td>“Abarrotes” (grocery); “Vende fruta preparada” (sell prepared fruit)</td>
</tr>
<tr>
<td>Mexico 2005</td>
<td>0.602</td>
<td>“Vendor”; “Tuckshop”</td>
</tr>
<tr>
<td>Kenya 2013</td>
<td>0.760</td>
<td>“Assorted goods”; “Grocery”</td>
</tr>
<tr>
<td>Zimbabwe 2017</td>
<td>0.649</td>
<td></td>
</tr>
</tbody>
</table>

This analysis aims to explore further the primary types of instances in which the more standard bag-of-words approach to measuring semantic difference differs from the sentence embeddings approach. Table A2 displays correlations between the two versions of the Differentiation Score across datasets and shows examples of “false positives”: businesses classified as highly differentiated using the bag-of-words approach but not highly differentiated with sentence embeddings.

A few themes emerge from this exercise: false positives tend to be either very generic businesses that happen to use less common terms, e.g. “general merchandise”, or those business descriptions in another language or with typos (as discussed in the main text, SBERT’s robustness to typos makes it a boon for survey data collected in the field).

Figures A2-A4 replicate the primary results using the bag-of-words-based differentiation
Figure A3: Overall Relationship Between Differentiation and Firm Revenue, Using Bag-of-Words Measure

Left panel displays estimated coefficients from OLS regressions of mean-adjusted monthly revenues on the Differentiation Score within each dataset. The random effects analysis estimates that a one-standard-deviation increase in the differentiation score is associated with an 15 percent increase in revenue relative to the mean. Each model controls for proprietor age, gender, years of education, and enterprise sector (retail, services, manufacturing, agriculture, or other). Right panel displays a binned scatterplot of the relationship between monthly revenues and the Differentiation Score within the combined dataset, controlling for the same covariates as in the left-hand panel and with an additional fixed effect for each individual study.

While overall similar, the results suggest that the bag-of-words approach may overstate the revenue and employment relationships while understating the profit relationship. In other words, it may be capturing business size more than a direct indicator of competitive performance. This may be an artifact of the fact that many of the “false positives” describe selling general merchandise of some kind, e.g., “assorted goods”, perhaps capturing the fact that these are larger variety shops, though not actually differentiated in any material sense apart from size.
Figure A4: Overall Relationship Between Differentiation and Firm Profit, Using Bag-of-Words Measure

Left panel displays estimated coefficients from OLS regressions of mean-adjusted monthly profit on the Differentiation Score within each dataset. The random effects analysis estimates that a one-standard-deviation increase in the differentiation score is associated with a 5 percent increase in profit relative to the mean. Each model controls for proprietor age, gender, years of education, and enterprise sector (retail, services, manufacturing, agriculture, or other). Right panel displays a binned scatterplot of the relationship between monthly profits and the Differentiation Score within the combined dataset, controlling for the same covariates as in the left-hand panel and with an additional fixed effect for each individual study.

<table>
<thead>
<tr>
<th>Study</th>
<th>Coef</th>
<th>SE</th>
<th>95%-CI Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philippines 2008</td>
<td>0.02</td>
<td>0.0614</td>
<td>0.02 [-0.10; 0.14] 16.9%</td>
</tr>
<tr>
<td>Uganda 2007</td>
<td>0.09</td>
<td>0.1133</td>
<td>0.09 [-0.14; 0.31] 11.9%</td>
</tr>
<tr>
<td>Togo 2013</td>
<td>-0.22</td>
<td>0.0701</td>
<td>-0.22 [-0.36; -0.08] 15.0%</td>
</tr>
<tr>
<td>Bangladesh 2010</td>
<td>0.08</td>
<td>0.0457</td>
<td>0.08 [-0.01; 0.17] 18.5%</td>
</tr>
<tr>
<td>Kenya 2013</td>
<td>0.16</td>
<td>0.0373</td>
<td>0.16 [0.11; 0.22] 20.0%</td>
</tr>
<tr>
<td>Zimbabwe 2017</td>
<td>0.14</td>
<td>0.0565</td>
<td>0.14 [0.02; 0.25] 17.2%</td>
</tr>
<tr>
<td>Random effects model</td>
<td></td>
<td></td>
<td>0.05 [-0.06; 0.16] 100.0%</td>
</tr>
</tbody>
</table>

Figure A5: Overall Relationship Between Differentiation and Firm Employment, Using Bag-of-Words Measure

Left panel displays estimated coefficients from OLS regressions of mean-adjusted total employment on the Differentiation Score within each dataset. The random effects analysis estimates that a one-standard-deviation increase in the differentiation score is associated with a 15 percent increase in employment relative to the mean. Each model controls for proprietor age, gender, years of education, and enterprise sector (retail, services, manufacturing, agriculture, or other). Right panel displays a binned scatterplot of the relationship between total employment and the Differentiation Score within the combined dataset, controlling for the same covariates as in the left-hand panel and with an additional fixed effect for each individual study.

<table>
<thead>
<tr>
<th>Study</th>
<th>Coef</th>
<th>SE</th>
<th>95%-CI Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philippines 2008</td>
<td>0.21</td>
<td>0.0634</td>
<td>0.21 [0.09; 0.34] 19.9%</td>
</tr>
<tr>
<td>Bangladesh 2010</td>
<td>0.01</td>
<td>0.0149</td>
<td>0.01 [-0.02; 0.04] 23.1%</td>
</tr>
<tr>
<td>Mexico 2005</td>
<td>-0.02</td>
<td>0.0385</td>
<td>-0.02 [-0.09; 0.06] 21.9%</td>
</tr>
<tr>
<td>Kenya 2013</td>
<td>0.30</td>
<td>0.0392</td>
<td>0.30 [0.22; 0.37] 21.9%</td>
</tr>
<tr>
<td>Zimbabwe 2017</td>
<td>0.34</td>
<td>0.1324</td>
<td>0.34 [0.08; 0.60] 13.3%</td>
</tr>
<tr>
<td>Random effects model</td>
<td></td>
<td></td>
<td>0.15 [0.01; 0.23] 100.0%</td>
</tr>
</tbody>
</table>

Employment (Mean Adjusted)

Differentiation Score (Standardized, Bag-of-Words)
What constitutes a ”Micro” Enterprise? Robustness from the Bangladesh Sample

The data from the Bangladesh sample (McKenzie, 2010) provide an opportunity to explore to what extent the methodology in this paper breaks down when applied to substantially larger firms. The original study surveyed randomly selected business owners in urban Bangladesh, with the criteria being that the businesses be informal (unregistered). The vast majority of these firms are indeed small – 90 percent have 10 or fewer employees and 80 percent have five or fewer – but the sample does include a long right tail of much larger firms, with a handful of firms with more than 100 employees. In the primary analysis, I drop firms with 15 or more employees, to be more in line with the other samples in the study. The following analysis shows the results replicated with the full sample of Bangladeshi firms, including the long tail of substantially larger firms.

Including the larger firms serves to increase the point estimate of the relationship between performance and differentiation, with a substantial loss in precision. A concern with applying the sentence embeddings method to measure differentiation in larger firms is that short descriptions are simply not sufficient to capture the more complex dimensions of competitive positioning in larger, more complex firms. The results from the Bangladesh sample suggest that while overall, the method may still serve to capture something relevant to performance, additional noise is introduced when larger firms are included. While the overall findings are unchanged, this exercise suggests that adaptations to the method described in this article may be warranted for larger firms (see Guzman and Li (2021) for a similar method applied to more complex firms).
**Figure A6:** Overall Relationship Between Differentiation and Revenue, With and Without Truncated Bangladesh Sample

Each panel displays estimated coefficients from OLS regressions of mean-adjusted monthly revenues on the Differentiation Score within each dataset, following the same conventions as in the primary analysis. Left panel caps the Bangladesh sample at firms with less than 15 employees, as in the main analysis. Right panel replicates the analysis with the full sample of firms from the Bangladesh study.

**Figure A7:** Overall Relationship Between Differentiation and Profit, With and Without Truncated Bangladesh Sample

Each panel displays estimated coefficients from OLS regressions of mean-adjusted monthly profits on the Differentiation Score within each dataset, following the same conventions as in the primary analysis. Left panel caps the Bangladesh sample at firms with less than 15 employees, as in the main analysis. Right panel replicates the analysis with the full sample of firms from the Bangladesh study.