van Angeren Joey (Orcid ID: 0000-0002-0501-6381)

Optimal distinctiveness across revenue models:

Performance effects of differentiation of paid and free products in a mobile app market

Joey van Angeren*
School of Business and Economics
Vrije Universiteit Amsterdam
de Boelelaan 1105, 1081 HV Amsterdam, the Netherlands
ORCID: 0000-0002-0501-6381
joey.van.angeren@vu.nl

Govert Vroom
IESE Business School
University of Navarra
Av. Pearson 21, 08034 Barcelona, Spain
ORCID: 0000-0002-7271-2723
gvroom@iese.edu

Brian T. McCann
Owen Graduate School of Management
Vanderbilt University
401 21st Avenue South, Nashville, TN 37203, USA
ORCID: 0000-0003-0251-5182
brian.mccann@owen.vanderbilt.edu

Ksenia Podoynitsyna
Jheronimus Academy of Data Science
Joint Institute of Tilburg University and Eindhoven University of Technology
Sint Janssingel 92, 5211 DA 's-Hertogenbosch, the Netherlands
ORCID: 0000-0003-1066-9165
k.s.podoynitsyna@tilburguniversity.edu

Fred Langerak
Department of Industrial Engineering and Innovation Sciences
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, the Netherlands
f.langerak@tue.nl

* Corresponding author

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/smj.3394

Running head: Optimal distinctiveness across revenue models

Keywords: differentiation, legitimacy, machine learning, optimal distinctiveness, revenue models

Optimal distinctiveness across revenue models:

Performance effects of differentiation of paid and free products in a mobile app market

RESEARCH SUMMARY

The optimal distinctiveness literature highlights a fundamental trade-off in product positioning within market categories: products should be distinct to minimize competition, but similar to build legitimacy. Most recently, this research has focused on understanding sources of variance in the distinctiveness-performance relationship. We extend this literature with an examination of digital products and argue that the relationship depends on products' revenue models: we theorize the relationship is inverted U-shaped for paid products but U-shaped for free products, owing to heightened privacy concerns of free product customers. We further argue that this latter relationship becomes flatter for free products that provide greater monetization transparency by publishing a privacy statement or adopting a freemium revenue approach. Hypotheses are tested using a sample of 250,000-plus Apple App Store apps.

MANAGERIAL SUMMARY

How should firms in the digital space position their products for optimal performance? We study this question in the Apple App Store, and suggest that the optimal positioning of digital products depends on their revenue model. Paid products should be moderately differentiated from competing products. By contrast, free products benefit most from very low or very high levels of differentiation. We attribute the different performance effects of differentiation to customers' privacy concerns over free products. Firms can partially ameliorate those privacy concerns by providing greater monetization transparency by publishing a privacy statement or by adopting a freemium revenue approach, making moderate levels of differentiation more viable. Our findings help managers align choices of positioning and revenue model, two critical aspects of the firm's business model.

INTRODUCTION

Research at the confluence of strategic management and organization theory highlights the competing pressures firms face when positioning products in market categories (Zhao *et al.*, 2017). On the one hand, products should be different from rival products in order to reduce competitive pressures (Cennamo & Santalo, 2013; Wang & Shaver, 2015). On the other hand, products benefit from being similar to rival products due to increased legitimacy (DiMaggio & Powell, 1983; Zuckerman, 1999). On net, a significant portion of this literature has recommended that firms balance these competing pressures and offer products that are moderately different in order to achieve "optimal distinctiveness" (Deephouse, 1999).

Consistent with this argument, a number of studies have found that a moderate level of distinctiveness is associated with the *highest* level of performance (e.g., Askin & Mauskapf, 2017; Deephouse, 1999); in other words, these studies suggest an inverted U-shaped relationship exists between degree of distinctiveness and performance. Other studies, in contrast, have produced different results, including those suggesting a U-shaped relationship where a moderate level of distinctiveness actually yields the *lowest* level of performance (e.g., Cennamo & Santalo, 2013; Miller *et al.*, 2018). This contrast in findings has inspired a growing effort by researchers to develop an understanding of factors that might explain why this relationship varies (e.g., Haans, 2019; Zhao *et al.*, 2018).

Most recently, a number of scholars have spotlighted the role of legitimacy in explaining variance in the distinctiveness-performance relationship, where legitimacy represents the degree to which customers understand a product and view it as meeting their expectations of what that type of product should do (Suchman, 1995). Legitimacy comes from a match between product characteristics and customer understanding and expectations. A legitimate product is one that a customer comprehends and views as being desirable and proper. In the context of explaining variance in the distinctiveness-performance relationship, Taeuscher and Rothe (2021) tie differences to the existence of alternative sources of legitimacy, while Taeuscher *et al.* (2021) focus on the role of differences in expectations across different audiences. This theoretical debate

around the role of legitimacy in determining optimal distinctiveness suggests there is more to understand about how customers assess legitimacy and how that may affect the shape of the relationship between distinctiveness and performance.

To extend this understanding, we explore how differences in products' revenue models affect customers' legitimacy assessments, leading to variance in the distinctiveness-performance relationship. We explain how legitimacy assessments vary because privacy concerns differ across different revenue models. A revenue model describes the monetization approach a firm pursues to generate sales from its products (Casadesus-Masanell & Zhu, 2010); it represents how a firm captures value from its products or services. Revenue models are one of the key aspects of a firm's overall business model (Massa *et al.*, 2017), and research highlights how revenue models may be a source of innovation (Snihur & Zott, 2020), resulting in varying approaches across firms. In particular, we note that the growing prominence of digital markets highlights a fundamental distinction in approaches, namely the distinction between paid and free revenue models (Eckhardt, 2016; Tidhar & Eisenhardt, 2020). Our central argument is that the shape of the distinctiveness-performance relationship varies across firms with paid versus free models, and we tie this variance to differences in legitimacy pressures across the two types of revenue models.

We begin with the base case of paid products and explain how we expect the operation of competitive and legitimation pressures to combine to produce an inverted U-shaped relationship between distinctiveness and performance, consistent with numerous prior studies in the literature (e.g., Askin & Mauskapf, 2017; Deephouse, 1999). Our theorizing then turns to the unique case of free products, and we explain that choosing increasingly distinctive market positions comes with a relatively higher legitimacy penalty for free products. Customers are naturally skeptical of products offered at no charge, especially in digital contexts where privacy is an increasingly important concern among consumers and perceived privacy risks can be substantial. This enhanced privacy concern leads to greater risk perception and more serious losses of legitimacy as the product deviates from the categorical prototype. The categorical prototype represents the typical or expected features and functionalities of products in a category (Durand & Paolella, 2013; Rosch

& Mervis, 1975). This leads us to predict an opposite relationship for free products, namely a U-shaped relationship between distinctiveness and performance. Finally, we deepen our theoretical contribution with a closer examination within the class of free products. Given the central role of privacy concerns in our prediction for free products, we theorize contextual factors that may attenuate these concerns resulting in moderation of the distinctiveness-performance relationship for free products. We contend that free products vary in their monetization transparency, or how easy it is for customers to understand the ways in which firms do and do not generate revenue from their products. First, some firms include posted privacy statements; these statements explicitly describe how the firm might use customer data to generate revenue, and these types of statements have been shown to decrease perceived privacy risks associated with transacting (e.g., Hui *et al.*, 2007). Second, firms offering free products may also increase transparency by following a "freemium" approach, where a base product is provided at no charge in concert with priced product upgrades. The freemium approach increases customer understanding of how the product will make money and thereby reduces concerns about potential misuse of customer information. In both cases, we contend that the increased transparency means products experience less legitimacy loss as they deviate from prototypes because customers are less suspicious of these products and perceive less privacy risk. As such, we predict a flattening of the U-shaped relationship for free products that adopt a privacy statement or a freemium approach.

We test our hypotheses in the context of the U.S. market of Apple's iOS App Store. With over two million distinct mobile apps and cumulative revenue in excess of $200 billion (Apple, 2020), the iOS App Store represents an economically significant digital platform where paid and free products engage in meaningful competition and in which positioning is an important determinant of product success. We observe the performance (number of downloads) of over 250,000 paid and free apps from four of the largest divisions of the App Store (Entertainment, Lifestyle, Productivity, and Utilities) over a period of eighteen months from May 2016 to October 2017. We use methods from machine learning, i.e., topic modeling and unsupervised Gaussian mixture model clustering, to identify market categories and characterize the positioning of individual apps.

Our empirical analyses generate results supportive of our theorizing.

We believe that our work offers several contributions. Product positioning is a fundamental competitive choice, and our research explains and demonstrates that the optimal choice systematically differs depending on whether a product is paid or free. Our research advances the optimal distinctiveness literature, which has most recently focused on explicating sources of variance in the distinctiveness-performance relationship. Work in this area has demonstrated that the distinctiveness-performance relationship varies across different categories (Haans, 2019), over time (Zhao *et al.*, 2018), across audiences (Taeuscher *et al.*, 2021), and with access to other sources of legitimacy (Taeuscher & Rothe, 2021). We extend this research by explaining how variance in a key aspect of a product's business model, namely its revenue model, has fundamental implications for how customers within a particular category view products; these efforts allow us to contribute to the optimal distinctiveness literature by further explicating the relevance and form that legitimacy concerns take. Whether a particular product meets the privacy expectations of a consumer can be a substantial driver of legitimacy, especially in the context of digital markets, leading to differences in the positioning-performance relationship. To the business model literature, we offer evidence of the importance of considering competitive and legitimacy pressures when designing business models, especially when determining a product's revenue model and positioning. Adopting a free revenue model is associated with greater legitimacy loss as distinctiveness increases; however, our work suggests that greater monetization transparency in the product's revenue model (e.g., through privacy statements and freemium) may reduce this negative effect. Our spotlight on the relationship of privacy concerns to legitimacy assessments also suggests the value of more deeply integrating privacy and cybersecurity issues into management scholarship. With over 80 percent of consumers saying the potential risks faced from data collection by companies outweigh the benefits (Pew Research Center, 2019), the degree to which this issue is effectively addressed by firms likely has substantial competitive implications. Finally, we also contribute to the platforms literature. We offer a method for finer-grained identification of market categories within platforms, extending prior literature relying on higher-

level categorization schemes (e.g., Barlow *et al.*, 2019; Eckhardt, 2016; Foerderer *et al.*, 2018). Our approach fosters a clearer focus on competitive strategy choices in typically densely populated digital platforms (Yoo *et al.*, 2012). Relatedly, we extend research examining intra-platform competition, which mostly focuses on the implications of platform providers' strategies for firms that produce complementary products (e.g., Foerderer *et al.*, 2018; Wen & Zhu, 2019), with our attention to the strategy decisions of those firms.

THEORETICAL BACKGROUND AND HYPOTHESIS DEVELOPMENT

Product positioning and optimal distinctiveness

Positioning is a long-established antecedent of a product's economic performance (Cennamo & Santalo, 2013; Porter, 1996; Wang & Shaver, 2015). Broadly defined, product positioning concerns a firm's choice of where to locate its product relative to competing ones within the boundaries of a category. The positioning dilemma is thus concerned with the question of how similar or distinct a product should be relative to other products in the same category. The positioning of a product is a consequence of a firm's choices concerning product attributes, notably product functionalities and features (Adner *et al.*, 2014). By implication, products are positioned farther from one another when their attributes are more different.

Academic interest in product positioning has reinvigorated with the emergence of the optimal distinctiveness literature (Zhao *et al.*, 2017). Scholars in this area conceive of product positioning as a balancing act between opposing demands from audiences to differentiate and conform in order to garner sales (Deephouse, 1999). The need for differentiation stems from competition in the category in which the product is set. Making a product more distinct allows it to stand out and be noticed by the members of an audience, thereby avoiding the most intense rivalry (Cennamo & Santalo, 2013; Wang & Shaver, 2015). Meanwhile, the need for conformity is rooted in the requirement for products to attain legitimacy in the eyes of an audience: "a generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate" (Suchman, 1995, p. 574). Institutional theorists conceptualize legitimacy as a multi-dimensional construct (Scott, 1995; Suchman, 1995). The cognitive legitimacy of a product depends on how comprehensible it

is to its audiences. Cognitive legitimacy facilitates product discovery, understanding, and comparison, and reduces the risk that audience members will come to question what the product does or why (DiMaggio & Powell, 1983; Zuckerman, 1999). In contrast, the normative legitimacy of a product depends on the degree to which it meets the expectations of its audiences. Unlike cognitive legitimacy, which depends on whether an audience understands what a product *actually* does, normative legitimacy relates to an assessment of whether a product does what it *should* do per audience expectations (Scott, 1995; Suchman, 1995).

Early consensus in the optimal distinctiveness literature was that there is an inverted U-shaped relationship between distinctiveness and economic performance (Zhao *et al.*, 2017);[1] the performance of a product first increases and then decreases with distinctiveness. The logic behind this understanding builds on the clustering tendencies of categories (DiMaggio & Powell, 1983; Zuckerman, 1999). Categories are generally characterized by a clear prototype that represents the most typical features and functionalities of the products in the category (Durand & Paolella, 2013; Rosch & Mervis, 1975). Distance from the prototype or category center first affects strength of competitive pressures. An indistinct product is similar to many other products in the category and faces substantial competition because it does not stand out. As the product becomes more distinct, the number of other products that it has to compete with for attention of audience members rapidly reduces. Additional efforts to make the product more distinct will do relatively little in further reducing competition because it essentially means that the product moves from one relatively uncontested position to another (Haans, 2019). Thus, distinctiveness first substantially reduces competition, but the reductions become less substantial as distinctiveness further increases.

Distinctiveness also influences legitimacy. A product that is very similar to what is considered the category's prototype is perceived as highly legitimate (DiMaggio & Powell, 1983). Owing to its close resemblance to the prototype, audience members are easily able to understand the product (cognitive legitimacy) and are presented with an offering that clearly aligns with their normative

[1] The optimal distinctiveness literature focuses on how distinctiveness affects the revenue aspects of profitability. It does not argue for any relationship between distinctiveness and cost.

expectations (normative legitimacy). As the product moves away from the prototype, legitimacy is lost as the lack of resemblance to the categorical prototype causes audience members to question what the product does and why (Zuckerman, 1999). Their lack of understanding of the product also increases perceived risk, causing audiences such as customers to actively shy away from the product in the presence of myriad alternatives given their risk-aversion (Pontikes, 2012), leading to reductions in sales. This is also true in a digital context. As just one recent example, Taeuscher and Rothe (2021) argue that the performance implications of positioning of Massive Open Online Courses depend on the legitimacy that such platforms possess from having high-status organizations as complementers.

Predicting the overall relationship of distinctiveness to performance requires combining the competitive and legitimacy pressures. That is, the shape of the distinctiveness-performance relationship depends on the relative strength of the two constituent pressures over the range of distinctiveness (Haans, 2019). Prior literature (e.g., Askin & Mauskapf, 2017; Deephouse, 1999) typically argues that the combination of competitive and legitimizing pressures results in lowest economic performance for products with very low and very high levels of distinctiveness but highest performance for those with moderate distinctiveness, leading to a net inverted U-shaped relationship between distinctiveness and performance.

Recently, the literature has begun to systematically investigate how this inverted U-shaped relationship might vary. In the context of Dutch creative industries, Haans (2019) shows that the shape of the relationship depends on whether products in a category are closely concentrated or have greater variance in their features and functions. Zhao *et al*. (2018) demonstrate that the age of a category exerts a substantial influence on the form of the distinctiveness-performance relationship while Taeuscher *et al*. (2021) show that the relationship varies depending on audience characteristics, specifically their taste for distinctiveness. Finally, Taeuscher and Rothe (2021) claim that the shape of the distinctiveness-performance relationship depends on the availability of other sources of legitimacy. In our work, we argue that differences in revenue models represent another fundamental distinguishing factor that has significant implications for the relationship of

product positioning and performance.

Revenue models and the distinction between paid and free products

A revenue model describes the monetization approach used to generate sales from products (Casadesus-Masanell & Zhu, 2010). Revenue models address how firms capture value, an essential aspect of their overall business model, where a business model is a specification of how the firm creates value, captures value through customer payments, and converts those payments to profits (Zott *et al.*, 2011). A variety of different types of business models have been described in the literature, including "discount," "razor-blade," and "sponsor-based" business models (Casadesus-Masanell & Zhu, 2010; Tripsas & Gavetti, 2000). What is notable about many of these different types of business models is how they are typically labeled in terms of their revenue model.

In the digital context, one particularly pertinent distinction is that between paid and free revenue models (Eckhardt, 2016; Ghose & Han, 2014; Kummer & Schulte, 2019; Tidhar & Eisenhardt, 2020). Mounting competition and zero-marginal costs of production and distribution for digital products have induced many firms to make their products available for free, possibly generating revenue from the products in some other way, rather than charging for them directly (Bryce *et al.*, 2011). In practice, this implies that in digital markets paid and free products coexist and compete in the same category (Ghose & Han, 2014). Indeed, in a study of software applications for Palm PDAs, Eckhardt (2016) showed that the performance of a paid application depends not only on the number of other paid applications in the same category, but also on the number of other free applications in this category. The same was shown to hold true for free applications.

A mounting body of literature in customer psychology suggests that customers' evaluations of paid and free products are markedly different (Hsu & Lin, 2015; Palmeira & Srivastava, 2013; Shampanier *et al.*, 2007). This research indicates that free products represent a distinct class in customers' minds, as they perceive them differently than alternatives of even negligible cost. In what follows, we will further theorize the consequences of this variance as it pertains to the distinctiveness-performance relationship for paid and free products. We home in on how differences in legitimacy pressures across the two types of revenue models result in differing

performance implications (i.e., quantity downloaded or sold).

The performance implications of distinctiveness for paid products

We first anticipate that the relationship of distinctiveness to performance for paid products follows an inverted U-shaped relationship, consistent with conventional arguments in the optimal distinctiveness literature (e.g., Askin & Mauskapf, 2017; Deephouse, 1999). In terms of competition, competitive pressures are strongest when there is little distinctiveness. Prior studies have illuminated this competition concern in the digital context, where prototypical products are frequently imitated by other firms for their success (Wang *et al.*, 2019). As paid products increase distinctiveness, competition falls quickly as differentiation is established from the prototypical product at the core of the category. Additional distinctiveness, however, provides marginally diminishing benefits. Once the product has largely broken free of the intense rivalry around the categorical prototype, added distinctiveness does little to further reduce competitive pressures.

In terms of legitimacy, we argue that paid products may exhibit small deviations from the categorical prototype without substantial legitimacy penalties. Customers will still be able to cognitively understand moderately distinctive paid products, and the products will maintain normative legitimacy owing to customers' flexibility in expectations concerning the incorporation of prototypical features (Deephouse, 1999). In fact, some degree of distinctiveness might actually be normatively desirable for paid products because customers generally expect some value for money (Hsu & Lin, 2015). However, as the product moves beyond this point, legitimacy is quickly lost.

In accordance with the above, Figure 1a illustrates how distinctiveness affects competition and legitimacy in the case of paid products. We follow the approach of Haans (2019) and Taeuscher and Rothe (2021) in depicting the constituent legitimacy and competition mechanisms as S-shaped curves.[2] In particular, Figure 1a shows how the level of competition drops before the level of

[2] Specifically, we follow their approach to depict the S-shaped curves as an inverse logit function ($\exp(b_0 + b_1X + b_2X^2)/(1+ (\exp(b_0 + b_1X + b_2X^2))$, where X represents distinctiveness. For the legitimacy curve in Figure 1a, b_0=4, b_1=-12, and b_2=0; for the competition curve, b_0=4, b_1=-10, and b_2=0.

legitimacy. Put differently, a moderately distinct paid product forgoes the most intense competition while retaining its legitimacy. By contrast, an indistinct paid product experiences substantially more competition but is hardly more legitimate, while a highly distinctive paid product foregoes the most intense competition but loses its legitimacy. This leads to our baseline prediction that moderately distinct paid products will exhibit the highest performance, yielding an inverted U-shaped relationship between distinctiveness and product performance.

--- Insert Figure 1a about here ---

Hypothesis 1. *The relationship between distinctiveness and performance is inverted U-shaped for paid products.*

The performance implications of distinctiveness for free products

It is our expectation that the shape and nature of the competitive pressures remain unchanged when customers evaluate free as opposed to paid products. Paid and free products are set in the same categories and are therefore confronted with similar competitive challenges (Eckhardt, 2016). However, we do expect legitimacy to more quickly erode with distinctiveness in the case of free products.

An important consideration, especially in the context of digital products, are privacy concerns often entailed when using free products. The emergence of free revenue models has prompted the rise of indirect monetization strategies, such as targeted advertisements or the selling of customer information, the success of which is critically contingent on amassing large volumes of user data (Casadesus-Masanell & Hervas-Drane, 2015; Goldfarb & Tucker, 2011). This naturally leads to skepticism and privacy concerns among customers, who question how a free product makes money and what happens to their personal data in the process (Martin *et al.*, 2017), and these concerns can have considerable impact on product appeal. For example, Al-Natour *et al.* (2020) demonstrate how privacy uncertainty affects the perceived risk associated with using mobile apps, reducing customers' intention to use. These concerns are particularly germane to free products given that in most settings it is rather hard for customers to observe the exact monetization strategies that free products employ as well as what data they collect (Hermalin & Katz, 2006). Paid products do not

fall under the same degree of scrutiny as their ability to generate revenues does not typically rely on the use of personal data. Indeed, in a study of mobile apps on Google Play, Kummer and Schulte (2019) show that the number of privacy-invasive permissions requested by free apps is much larger than the number requested by paid apps.

We contend that heightened privacy risk aversion surrounding free products affects the relationship between legitimacy and distinctiveness. More specifically, we anticipate both cognitve and normative legitimacy are harder to retain as distinctiveness increases. Customers are better able to gauge the risks associated with a product when it closely resembles the categorical prototype simply because it is easier for customers to understand them. But, when products become a little more distinctive and therefore somewhat more difficult to understand, customers will quickly question what a free product does, why it does it, and how this affects them given their privacy concerns. This results in a quicker drop in legitimacy, relative to the case of paid products, as distinctiveness increases. Stated another way, while indistinctive free products will be perceived as legitimate, legitimacy will rapidly be lost once a free product becomes more distinctive.

Figure 1b illustrates the theorized shift in the legitimacy-performance relationship for free products compared to paid products, highlighting the quicker drop in legitimacy as the product deviates from the prototype in the category.[3] The revised legitimacy curve of free products is joined with the unchanged competition curve in Figure 1c, which shows that, in contrast to the situation for paid products, legitimacy now drops before competition. And, as Haans (2019) describes, whether the legitimacy or competition curve drops first determines the shape of the distinctiveness-performance relationship. As such, our theorized shift in the legitimacy curve for free products results in a U-shaped distinctiveness-performance relationship. This relationship indicates that moderately distinctive free products will exhibit the worst, rather than the best performance.

--- Insert Figure 1b and 1c about here ---

Hypothesis 2. *The relationship between distinctiveness and performance is U-shaped for free*

[3] Beta coefficients for the legitimacy curve for free products: b_0=6, b_1=-12, and b_2=0.

products.

The moderating role of monetization transparency for free products

We suggest that customer concerns about how their personal information will or will not be used and the associated legitimacy losses will be attenuated in cases where a free product provides greater transparency about its revenue model. We use the term *monetization transparency* to refer to the degree to which customers may easily understand how the product does and does not generate revenue for the firm. Transparency involves disclosure of information about the practices, policies, and procedures of an organization. Prior research suggests that transparency plays a role in developing and maintaining legitimacy. For example, Gegenhuber and Dobusch (2017) describe how new ventures gain legitimacy by adopting an "open strategy" approach that provides high levels of transparency. Luedicke *et al.* (2017) also argue that the transparency of the radically open strategizing process of the German Premium-Cola collective was a key contributor to building organizational legitimacy. Desai (2018) offers similar arguments with the contention that increased transparency that accompanies open, collaborative engagements with external stakeholders is a key contributor to legitimacy. In sum, transparency may increase customer confidence that the firm's actions are appropriate, proper, and desirable (i.e., legitimate) (Suchman, 1995). Monetization transparency involves clear disclosure of information about the firm's policies related to the revenue model of its product, and we anticipate that this transparency can be achieved through both publishing privacy statements and utilizing a freemium approach.

Privacy statements. As customer apprehension is directly associated with privacy concerns, a straightforward way to address these concerns is to provide an explicit statement or policy describing the firm's privacy practices.[4] Such statements can help customers better understand what the product provider will and will not do with their data. In essence, the privacy policy provides explicit detail to support the implied social contract underlying the exchange of personal information for the use of a product (Dunfee *et al.*, 1999). The social contract perspective suggests

[4] A privacy policy provides a detailed explanation of a firm's privacy practices while a privacy statement is a relatively short, accessible summary that usually also directs customers' attention to the more detailed privacy policy.

that customers will consider provider privacy policies before transacting, and the level of trust they place in the provider (and hence their willingness to transact) is influenced by a provider's use of trustworthy privacy practices. Research extending back to the early stages of e-commerce argues that customer perceptions of privacy risk will be diminished when their concerns about privacy are addressed by fair procedures, and one key aspect of these fair procedures is disclosure of privacy practices (Culnan & Armstrong, 1999).

A wealth of empirical work has examined the effect of privacy statements, and these studies generally indicate their benefits in reducing privacy risk concerns. Pan and Zinkhan (2006) found more favorable customer responses to online shopping sites that included clearly stated privacy messages compared to those without a statement. This effect was particularly strong for customers with heightened concerns of privacy risk. Hui *et al.'s* (2007) field experiment indicated that customers were more willing to disclose personal information to web sites that included a privacy statement. Miyazaki (2008) showed that customer detection of cookie use by a website led to a decrease in customer trust and decreased purchase intentions; however, these negative effects were attenuated by the use of disclosure statements regarding the use of cookies. It is also important to note that privacy statements are not mere "cheap talk" (Farrell & Rabin, 1996) because the content of the policies may create legal liability for product providers (Hintze, 2018).

In sum, we anticipate that the increased transparency associated with the inclusion of privacy statements reduces privacy concerns. As such, products that include these statements experience less legitimacy loss when positioned away from category prototypes compared to those without privacy statements. In other words, the legitimacy curve for free products with privacy statements does not drop as quickly compared to products without privacy statements.[5] The slower dropping legitimacy curve suggests that the U-shaped distinctiveness-performance relationship will be flattened for free products providing an explicit privacy statement.

Hypothesis 3. *Use of a privacy statement moderates the U-shaped relationship between distinctiveness and performance for free products such that the U-shape is flatter for products with a privacy statement.*

[5] Graphically, this would be represented by a legitimacy curve in between the two curves depicted in Figure 1b.

Freemium. Besides explicit policy disclosures, we also anticipate that firms may increase transparency via other choices. We focus here on free products that utilize "freemium" revenue approaches. As Rietveld (2018) notes, freemium offerings have become increasingly ubiquitous in digital goods, such as video games, mobile apps, and social networking services. A free product with a premium option (i.e., freemium) allows customers to access basic features of a product or service for free and then charges for upgrades to this basic version. The video conferencing software *Zoom*, for example, offers a free version that lets users host a video conference call constrained to a limited duration for a limited number of participants. Paid versions permit more participants and longer calls. The mobile video game *Clash of Clans* offers a variety of upgrades and add-on features that users can purchase to augment the free version.

What matters for the purposes of our theorization is that freemium offerings provide increased monetization transparency. Customers are better able to understand how a firm generates revenue from its products. They realize that a free product is offered as an enticement to try the product and perhaps upgrade by paying for enhanced features, which obviously provides sales revenue to the firm. Correspondingly, customers' concerns related to potential invasive monetization strategies, such as the sale of customer information to third parties, are reduced. These reduced concerns associated with increased transparency mean that freemium products experience less legitimacy loss when they position themselves away from category prototypes. In contrast, free products without the monetization transparency associated with a freemium offering suffer more substantial legitimacy penalties when departing from category norms. In other words, the legitimacy curve of freemium products does not drop as quickly compared to free products not utilizing the freemium approach. These differences in the underlying relationship between distinctiveness and legitimacy suggest the overall distinctiveness-performance U-shape will be flattened for free products adopting the freemium approach.

Hypothesis 4. *Use of a freemium revenue approach moderates the U-shaped relationship between distinctiveness and performance for free products such that the U-shape is flatter for products utilizing freemium.*

METHODS AND DATA

Empirical context and data

We tested our hypotheses in a sample of mobile apps from the U.S. market of Apple's iOS App Store. Mobile apps are small software applications that complement the standard functionality of customers' mobile devices. Introduced in 2008, the iOS App Store is an economically significant example of a digital platform. It contains more than 2.2 million distinct apps that have generated more than $200 billion in cumulative revenue (Apple, 2020). Apple screens apps to ensure they meet a minimum legitimacy threshold, refusing apps with obviously malicious purposes. Importantly, the iOS App Store also constitutes a context where paid and free apps coexist and compete for the attention of the same customers (Ghose & Han, 2014; Tidhar & Eisenhardt, 2020). Apps' distinctiveness is also measurable. App developers include a textual description that provides an overview of the app's functionalities and constitutes the main avenue through which a developer communicates its characteristics to potential customers (Barlow *et al.*, 2019).

We collected monthly observations on all apps in the app store categories Entertainment, Lifestyle, Productivity, and Utilities between May 2016 and October 2017 using web scraping. For expositional clarity, we will refer to these as "divisions" of the app store to avoid confusion with the "categories" we examine in the paper. We focused on those divisions because they are among the largest in the iOS App Store, yet comparably less likely to contain apps that merely serve as sales channels. We excluded apps whose descriptions were not in the English language and shorter than ten words because we rely on the descriptions for variable operationalization. We also excluded apps by developers with more than 500 apps as these are typically contract developers who develop apps for other companies and/or simply introduce the same white-labeled app many times; as such, the descriptions of these apps may not adequately reflect the positioning of the apps.[6] This left us with a sample of 268,126 apps: 72,017 paid and 196,019 free. We

[6] The substantive conclusions of our analyses do not depend on the decision of whether to include or exclude these observations. We also note that, although our data do not allow us to specifically identify them, our sample likely includes individuals who are hobbyist app developers whose motivation may not necessarily be to maximize performance. This is not an issue to our analyses, as we just need to observe *where* an app was positioned, not *why* it

collected rich app information beyond the description, including the title, developer name, release date, price, list of in-app purchase items, submitted customer ratings, and whether the app appeared on one of the sales leaderboards in the iOS App Store. We complemented this data set with proprietary data on periodic app performance by leading mobile app analytics company Apptopia.[7] Among other performance metrics, Apptopia provides data on the number of downloads.

Modeling app characteristics and categories through machine learning

The app store divisions in our sample contain anywhere between 40,000 and 110,000 apps. The smallest app store division, Productivity, contains apps ranging from QR code scanners, to contact backup services, and onto mind mapping tools. Many customers in the iOS App Store discover apps not by browsing an app store division, but by means of an organic search process (TUNE 2015). For example, a customer looking for a task management application might directly search for "task manager" in the iOS App Store, to then only evaluate the list of apps whose description somehow corresponds to this search criterion. By implication, these lists of apps constitute the categories in our setting, which are nested in the divisions in the iOS App Store.

Because those categories are not directly observable from our data, we utilized machine learning methods to first model the dimensions along which apps can be distinguished and to subsequently identify categories of apps. Following recent work on optimal distinctiveness (Haans, 2019; Taeuscher *et al.*, 2021), we applied Latent Dirichlet Allocation (LDA) to all preprocessed[8] descriptions to characterize apps (Blei *et al.*, 2003). LDA is a probabilistic model that uncovers topics that are latent in a collection of descriptions based on the observed co-occurrence of words.

---

was positioned there. Moreover, consumers are generally unaware of the exact identity of the developer, meaning their privacy concerns unlikely differ depending on developer identity.

[7] Apptopia (https://www.apptopia.com) is a leading mobile analytics company that provides a data and decision-making platform with insights on the mobile app industry. Its customers range from small developers to large prominent companies, such as Bloomberg, Facebook, Google, Lyft, and Visa.

[8] We performed some preprocessing before applying topic modeling. We removed numbers, punctuation, special characters, stop words (e.g., "and," "or," "app"), non-English words, words that occur in more than 50% of all descriptions per app store division, and words that occur in less than 10 descriptions per app store division. We used a lemmatization algorithm to convert all words to their root form (e.g., "performs," "performed," and "performing" becomes "perform") to reduce lexical complexity. We applied part-of-speech tagging to retain only nouns and verbs, i.e., the words that are most descriptive concerning apps' features. All those procedures were implemented in Python, using the Natural Language Toolkit (https://www.nltk.org).

In our case, we conceive of topics as representing meaningful dimensions that distinguish apps, most notably representing product functionalities and features. For example, a topic with keywords such as "push," "notification," and "alert" relates to a feature of sending users real-time updates. As such, LDA makes it possible to characterize each app by representing its description as a probabilistic distribution over topics, where the probabilities sum to one.

We applied LDA per app store division. Our LDA implementation is based on the online variational Bayes algorithm that learns topics by iteratively going through batches of descriptions (Hoffman *et al.*, 2010).[9] We used 200 learning iterations, using standard parameters for κ and τ_0, which control the learning rate and extent to which learnings from early iterations are devalued, respectively. We set the number of topics to 150 per app store division, rather than the oft-applied value of 100 topics (Haans, 2019; Taeuscher *et al.*, 2021), to be able to identify a greater depth of dimensions along which apps are being distinguished. Based on manual inspection of all topics and their most characteristic keywords, we removed between fifteen and twenty topics per app store division that did not capture meaningful dimensions of apps, and then rescaled apps' probabilistic topic distributions such that they again sum to one.[10] For example, from the Entertainment app store division we removed a topic with keywords "run," "background," "life," and "battery" representing a reminder for customers to close the app after using it to preserve the battery life of their mobile device. Similarly, from Productivity we omitted a topic characterized by the combination of keywords "please," "rate," and "review" that merely captured a developer's efforts to get customers to write a review for its app.

Our topic models capture many meaningful dimensions along which apps can be distinguished. Some topics are clearly centered on a dimension that might constitute a prototypical feature of a category, such as the topics containing "password," "generate," and "manager" or "code," "qr," and "scan" in Utilities. Other topics instead capture dimensions that correspond with less characteristic, supplementary features. For example, the topic with keywords "facebook,"

[9] We used the scikit-learn machine learning toolkit (https://www.scikit-learn.org).
[10] Our results are robust to skipping this step and retaining all topics.

"twitter," and "share" in Lifestyle, or the Entertainment topic containing "apple," "watch," "glance," and "wrist." Tables S1-S4 in the Online Appendix provide an overview of all topics per app store division based on their ten most characteristic keywords.

We identified app categories by clustering the apps in each app store division based on their topic distributions. The logic here is that we segregate the app store division into categories of apps that are similar enough to one another to appear jointly under the same search terms such that customers typically evaluate them together, yet distinctively different from apps in other categories. We inferred categories using the Gaussian mixture model clustering algorithm for its ability to identity categories that differ in size, shape, and density (McLachlan & Basford, 1988). This is important in our setting, because the iOS App Store simultaneously harbors general apps, such as hundreds of to-do list applications, alongside specific apps, such as a handful of drawing tools for architects. We allowed the number of categories per app store division to be borne out of the data by iteratively optimizing the Bayesian Information Criterion (BIC) (Schwartz, 1978). We identified 893 categories in Entertainment, 1,018 categories in Lifestyle, 465 categories in Productivity, and 712 categories in Utilities. The number of apps per category varies between 9 and 1,309, with an average of 127. Table S5 in the Online Appendix illustrates the mapping between app descriptions, topics, and categories.

Dependent variable

We measured app performance as the monthly number of *downloads*, using the proprietary information from Apptopia. The number of downloads is a suitable measure for app performance that directly reflects the app's legitimacy with customers and that can be applied consistently across both paid and free apps. Moreover, downloads directly relate to the revenues of paid apps, and are an important precursor to the successful monetization of free apps as they directly affect the stock of personal data available to the developer of the app. We used the natural logarithm of the number of downloads (plus 1) to reduce the skewness in this variable.

Independent and moderating variables

We followed Haans (2019) and measured *distinctiveness* as the difference between an app's topic

distribution and the average topic distribution for all apps in the same category.[11] That is:

$$distinctiveness_{it} = \sum_{T=1}^{N} abs(\Theta_{Tit} - \overline{\Theta}_{Tct})$$

where Θ_{Tit} refers to app *i's* weight for topic *T* in month *t* and $\overline{\Theta}_{Tct}$ is the average weight for topic *T* in category *c* in month *t*. We computed this measure at each month to reduce measurement error, accounting for the fact that the prototypical representation of a category can change over time (Zhao *et al.,* 2018) and that apps can be repositioned (Wang & Shaver, 2015), although repositioning occurs somewhat rarely in our context.

In the case of free apps, we also focused on *privacy statement* and *freemium* as moderating variables. Privacy statement was coded as a dummy variable that captures whether the app description contained an explicit privacy notice that elucidates how and why the app uses customer data. For example, the app description of *Mobref*, an app for small business owners to manage referrals from the Utilities app store division contains the following passage:

> *"Why does Mobref need my phone number and address book? Mobref requires your phone number to uniquely identify and deliver messages back and forth. We use metadata in your address book to identify businesses and services and enrich your "My businesses" section. Your address book data is private to you. No one using Mobref would gain access to your address book data, unless you share contacts with them. We value your privacy. You can read up about our Privacy Policy in detail."*

We coded for the presence of a privacy statement by searching for the presence of key phrases such as "data privacy," "data protection," "privacy policy," "privacy statement," "respect your privacy," and "your privacy is important" in the app description that are associated with statements about customers' privacy. Tables S6 and S7 in the Online Appendix contain the full list of key phrases and further examples of privacy statements from our data, respectively. We coded the privacy statement variable "1" if one or more of those key phrases was present in the app description, and zero otherwise.

Freemium was operationalized as a dummy variable that indicates whether a free app uses freemium. In our empirical context, free apps can operate freemium in two different ways. One, an app can offer a menu of upgrades available for in-app purchase, which we could readily observe

[11] We note one small distinction in our measure relative to Haans (2019), namely that our measure is time varying.

from our data. Two, a developer may associate a free app with a premium version made available as a separate paid app. We relied on a simple text matching procedure to match such pairs of free and paid apps. We applied subset detection to check whether free and paid apps shared part of the same title, and used cosine similarity, the normalized angle between two word vectors (Hoberg & Philips, 2016), to determine whether their descriptions were also by and large similar. Correspondingly, the freemium dummy variable was coded "1" if a free app had in-app purchases or when it was associated with a paid app, and is zero otherwise.

Control variables

We included several control variables at the level of the category. We controlled for the extent to which apps are spread out across a category, *category heterogeneity*, measured as the standard deviation of distinctiveness across all apps in the category, because it may affect the relative benefits of distinctiveness (Haans, 2019). Since the effect of distinctiveness may change with *category maturity* (Zhao *et al.*, 2018), we controlled for the number of months since the oldest app in the category was released. We factored in the effect of *category size*, by controlling for the log-transformed number of apps in the category. The percentage of paid apps in the category, *percentage paid in category,* might influence performance implications from distinctiveness, as paid and free apps might experience different competition and/or legitimacy pressures from rival paid apps as opposed to rival free apps (Eckhardt, 2016). We used the Herfindahl-Hirschman Index, *HHI*, defined as the sum of squared download market shares, to control for competition in the category. We also included dummy variables per app store division.

In addition, we controlled for numerous factors at the app and developer levels that might impact app performance (Eckhardt, 2016; Ghose & Han, 2014). We controlled for *developer category experience* through a log-transformed count of all apps by the developer in the category. We accounted for the amount of information that was available about an app, by controlling for *description length*, the number of words in the app description, and the number of *screenshots* that were displayed on the app's information page. *App age* captured the number of months since an app has been released. *File size* measured the log-transformed file size in megabytes. For paid

apps, we also include the log-transformed *price* in U.S. dollars. To control for app quality, all models included the log-transformed number of times an app had been rated (*ratings*) and a series of dummy variables representing distinct app rating levels between one and five stars. We accounted for the potential beneficial effects of visibility in the iOS App Store by controlling for the log-transformed number of *recommendations* of a focal app on the information pages of other apps, whether apps were *ranked* on one of the iOS App Store sales leaderboards (i.e., top free apps, top paid apps, and top grossing apps), and whether they were *featured* on one of the landing pages in the App Store (e.g., "Essentials," "Editor's Choice," or "Apps We Love"). A series of dummy variables controlled for apps' content rating. All models also included monthly dummies.

Estimation approach

We performed our estimations using a generalized estimation equation (GEE) panel data model that produces population average results (Liang & Zeger, 1986). Because apps are rarely repositioned, standard fixed effects models were inappropriate. We created two sets of GEE panel data regressions, one for paid apps and one for free apps. A test for auto-correlation indicated significant first-order serial correlation in our data. Therefore, we included a first-order autocorrelation (i.e., AR1) correlation matrix in our regressions. All our estimations were performed using robust standard errors.

RESULTS

The descriptive statistics and correlations for paid apps are in Table 1a; those for free apps are in Table 1b. As expected, free apps are downloaded more frequently than paid apps. Before log-transformation, the average number of monthly downloads for free apps is 759 (SD = 15,691), while the average paid app is downloaded 33 times a month (SD = 715).[12] We further observe that roughly 2.5% of the free apps in our sample have a privacy statement; 14% utilize freemium.

--- Insert Table 1a and Table 1b about here ---

[12] We note that while free apps are downloaded more frequently than paid apps, the average log-transformed number of downloads is actually lower for free apps compared to paid apps. This is because of the greater degree of skewness and zero-inflation in the distribution of downloads for free apps.

The estimation results for paid apps are in Table 2. Results of the baseline model are reported under Model 2.1. The control variables largely act in the expected manner. For example, app performance increases as apps are more heterogeneous across the category, as an app garners more ratings, and when an app is featured more prominently in the iOS App Store. App performance decreases with competition (HHI) in the category, app age, and app price. Model 2.2 introduces distinctiveness, and Model 2.3 introduces distinctiveness squared.

--- Insert Table 2 about here ---

We turn to Model 2.3 to test Hypothesis 1, in which we predicted an inverted U-shaped relationship between distinctiveness and performance for paid apps. Following Lind and Mehlum (2010) and Haans *et al*. (2016), we first look at the coefficients for distinctiveness and distinctiveness squared. We observe that the main effect of distinctiveness is positive and significant (b = 0.493, SE = 0.066, $p < 0.001$) and that the coefficient for distinctiveness squared is negative and significant (b = -0.191, SE = 0.034, $p < 0.001$), suggestive of an inverted U-shaped relationship. Second, we evaluate the slopes at the lowest (0) and highest (1.92) value of distinctiveness and find that the slope of distinctiveness is positive and significant (b = 0.493, SE = 0.066, $p < 0.001$) at the lowest level and negative and significant (b = -0.242, SE = 0.068, $p < 0.001$) at the highest level. Third, we establish the turning point of the curve, the point at which distinctiveness yields the highest level of app performance, which is at a distinctiveness value of 1.289, with a 95% confidence interval ranging between 1.114 and 1.438, well within the range of distinctiveness values in our data. We thus conclude that app performance first increases and then decreases with distinctiveness and therefore report support for Hypothesis 1. The inverted U-shaped distinctiveness-performance relationship for paid apps is visualized in Figure 2.

--- Insert Figure 2 about here ---

Table 3 presents the estimation results for free apps. Model 3.1 constitutes the baseline model, where we observe that the control variables largely have the anticipated effects. We add distinctiveness and distinctiveness squared in Model 3.2 and 3.3, respectively. To test Hypothesis 2, which predicted a U-shaped relationship between distinctiveness and performance for free apps,

we follow the same procedure as when testing Hypothesis 1. First, looking at Model 3.3, we observe that the coefficient for distinctiveness is negative and significant (b = -0.167, SE = 0.041, $p < 0.001$) and that the coefficient for distinctiveness squared is positive and significant (b = 0.063, SE = 0.021, p = 0.004), indicative of a U-shaped relationship between distinctiveness and performance. Second, we test the slopes at both ends of the distinctiveness distribution. The slope at the lowest level of distinctiveness (0) is negative and significant (b = -0.167, SE = 0.041, $p <$ 0.001). It is positive and marginally significant (b = 0.074, SE = 0.045, p = 0.098) at the highest level (1.92) of distinctiveness, suggesting that the distinctiveness-performance relationship is somewhat upward sloping at high levels of distinctiveness. Third, we determine that the turning point of the curve, where the returns to distinctiveness are lowest, lies at a distinctiveness value of 1.330 (95% confidence interval: [0.997, 1.663]). Taken together, these findings are consistent with a U-shaped distinctiveness-performance relationship for free apps, providing support for Hypothesis 2. Figure 3a visualizes this relationship.

--- Insert Table 3 and Figures 3a-3c about here ---

Model 3.4 and 3.5 introduce the interactions of privacy statement and freemium with distinctiveness and distinctiveness squared to test Hypotheses 3 and 4. Model 3.6 is the full model. Hypothesis 3 predicts a flattening of the U-shaped relationship between distinctiveness and performance for free apps with a privacy statement. Model 3.4 provides support for this hypothesis, as does Model 3.6. The results show that the interaction between distinctiveness squared and privacy statement is negative and significant (b = -0.613, SE = 0.211, p = 0.004), which suggests a flattening of the U-shaped relationship for free apps with a privacy statement (Haans *et al.*, 2016). We visualize the distinctiveness-performance relationship for free apps with and without a privacy statement in Figure 3b. Interestingly, it seems that the distinctiveness-performance relationship for free apps with a privacy statement flips and more closely resembles an inverted U-shape. We examine this more formally using the procedure outlined by Haans et al. (2016, p. 1188), and find that a shape flip indeed occurs, such that the relationship between distinctiveness and performance is no longer U-shaped for free apps with a privacy statement.

Hypothesis 4 stated that the U-shaped distinctiveness-performance relationship flattens for free apps using freemium. The coefficient for the interaction between distinctiveness squared and freemium is negative and significant in Model 3.5 (b = -0.207, SE = 0.068, p = 0.002), so Hypothesis 4 is also supported. Here too, a shape flip occurs. The distinctiveness-performance relationship for free apps that are freemium is no longer U-shaped, as shown in Figure 3c. Model 3.6 provides similar results.

Robustness checks

To probe the robustness of our findings, we began with addressing two potential sources of bias in our main analyses. First, because we used a split sample approach to analyze the distinctiveness-performance relationship across paid and free apps, and selection into either sample is not random, there is the potential for sample selection bias. Second, our variables of interest concern choices by developers giving rise to potential endogeneity concerns due to omitted variables.

We addressed the potential sample selection bias (paid versus free) by applying a Heckman correction and generating an inverse Mills ratio that we then included in our other regressions. We predicted whether an app is free or paid by means of a first-stage probit model, using the share of free apps in the portfolio of a developer (excluding the focal app) as an instrument. Regarding relevance, an increase in this variable increases the relative likelihood that the focal app will be free as well; prior work in our empirical context (Kummer & Schulte, 2020; Wang *et al.*, 2019) has emphasized how developer averages are good predictors of choices related to individual apps. Turning to exogeneity, the share of free apps in the portfolio of a developer has no clear theoretical relationship to the downloads of a focal app. Most developers in our context are small and have unknown reputations to customers (Arora *et al.*, 2017). Correspondingly, customers are familiar with and evaluate individual apps, but typically not developers. We also included a dummy variable for developers that had no other apps in their portfolio.

To deal with endogeneity concerns over the distinctiveness, privacy statement, and freemium variables, we applied a control function approach (Wooldridge, 2010), making use of comparable developer-side instrumental variables as prior work in our empirical context (Kummer & Schulte,

2020; Wang *et al.*, 2019). We ran first-stage models to predict distinctiveness for paid apps and free apps, the use of a privacy statement, and freemium for free apps and then included the residuals from those regressions as additional control variables in our main models, representing the component of the distinctiveness, privacy statement, and freemium variables that likely correlate with the error term. All models were estimated based on 1,000 bootstrap replications (Wooldridge, 2010). To predict distinctiveness, we used the average distinctiveness of the other apps in the portfolio of the developer as an instrument. Similar to the point mentioned above, the idea behind this instrument is that it captures a developer's inclination to differentiate its apps (relevance), but this inclination only affects downloads of the focal app through the level of distinctiveness of the focal app (exogeneity). To predict the use of privacy statement and freemium we followed a similar logic: we used the share of free apps with a privacy statement or that were freemium by a developer as an instrument for privacy statement and freemium, respectively. We also included the no-other-apps dummy variable in all first-stage models.

Tables S8, S9, and S10 in the Online Appendix present the results of all stages of the bias correction procedures. In contrast to two-stage least squares (2SLS) models, formal tests of relevance and exogeneity are unavailable under a control function approach. However, concerning the relevance of our instruments we note that they are strong predictors of the endogenous variables and work in the way we anticipated. Table S8 shows that share of free apps strongly predicts selection into paid versus free ($b = 2.155$, SE = 0.003, $p < 0.001$). Table S9 shows that average distinctiveness of other developer apps strongly predicts distinctiveness of the focal app for paid products ($b = 0.379$, SE = 0.005, $p < 0.001$). Table S10 shows that our instruments in the free sample are similarly strong (average distinctiveness of other apps: $b = 0.442$, SE = 0.004, $p < 0.001$; share of free apps with a privacy statement: $b = 3.398$, SE = 0.041, $p < 0.001$; share of free apps that are freemium: $b = 2.381$, SE = 0.017, $p < 0.001$). Moreover, prior work in our empirical context has provided evidence using a 2SLS framework of the exogeneity of comparable instruments for app developers' strategic choices, including privacy-invasiveness (Kummer & Schulte, 2019, p. 3479) and pricing (Wang *et al.*, 2019, p. 287-288). Applying sample selection

and endogeneity corrections, all our results hold.

We also performed a few other robustness checks. Our theoretical explanations hinge on the existence of a sharp discontinuity in how customers evaluate free as opposed to paid products. To evaluate the existence of this discontinuity in our data more formally, we performed a subsample analysis in the sample of paid apps. Specifically, we exclusively focused on apps priced at $0.99, which is the lowest pricing point in the iOS App Store, to verify that the characteristic inverted U-shaped distinctiveness-performance relationship for paid apps holds when apps are nearly free. We indeed produce an inverted U-shaped distinctiveness-performance relationship, providing further support for the idea customers evaluate free and paid products differently.

We ensured that our results are not affected by the way we operationalize app performance. We measured app performance as the log-transformed number of downloads. In a robustness check, we instead used apps' log-transformed number of daily active users (DAU) as the dependent variable. This measure captures intensive app usage by counting the number of unique users that open the app at least once a day. App usage is a critical indicator of performance in the mobile app industry (Chen *et al.*, 2021). The estimation results are equivalent to our main models.

A number of mobile apps in the iOS App Store are no longer actively maintained, and one might question whether these apps should be included in the sample as market participants. To verify that such deserted apps are not driving our results, we repeated our estimations using a sample of apps for which we observed one or more updates during our sampling period, leaving us with roughly 25% of the original sample for paid apps and 35% of the original sample for free apps. We obtain results identical to our main models in terms of direction, significance, and magnitude. Because inactive apps may no longer be perceived as legitimate competitors in the eyes of other developers and customers, we also recomputed our distinctiveness measure along with all other category-related control variables to only reflect an app's position relative to active competitors, dropping observations on apps without active competitors. Our findings hold.

We also reran our models using a different distinctiveness measure. Because customers may evaluate apps based on the presence or absence of certain characteristics, rather than based on the

relative prominence of those characteristics, we dichotomized[13] apps' topic distributions and then computed distinctiveness as the average pairwise normalized Hamming (1950) distance across all apps in the same category. The results remain unchanged when replacing our original distinctiveness measure with this alternative distinctiveness measure.

DISCUSSION

Firms face competing tensions in selecting the product attributes, functionalities and features that define the positioning of these products in market categories. On the one hand, products should conform to other products in the category in order to gain legitimacy. On the other hand, products should be different from other products in the category in order to avoid competitive pressure. How firms should manage these tensions in their quest for superior performance has been a continuing question of interest (Zhao *et al.*, 2017). While early research proposed that an intermediate level of distinctiveness delivered the highest level of economic performance (Askin & Mauskapf, 2017; Deephouse, 1999), more recent research has begun to challenge and explore the ubiquity of that result. More specifically, recent literature focuses on the question of under what conditions moderate distinctiveness is an optimal approach (e.g., Haans, 2019; Taeuscher *et al.*, 2021; Taeuscher & Rothe, 2021; Zhao *et al.*, 2018). The intent of this study was to further extend this important line of inquiry by investigating how the distinctiveness-performance relationship differs across products with different revenue models.

Our analyses used machine learning, based on topic modeling and Gaussian mixture model clustering, to operationalize the categories and distinctiveness of over 250,000 mobile apps from the Apple iOS App Store over an eighteen-month period. Importantly, our approach allowed us to establish a fine-grained classification of competing products on this large digital platform. Our results supported an inverted U-shaped relationship between distinctiveness and economic performance for paid products, suggesting that a moderate level of distinctiveness delivered the highest level of performance. However, we found the opposite result for free products, with high

[13] We designated a topic present in apps' probabilistic topic distributions if the corresponding topic loading is greater or equal than 0.05.

and low levels of distinctiveness outperforming moderate distinctiveness. Furthermore, this U-shaped relationship for free products flattened for products both with a privacy statement and using a freemium revenue approach. Indeed, our depiction of these results showed how the shape not only flattened, but actually flipped. Overall, our analyses clearly indicate that variance in products' revenue models is a critical factor that affects the distinctiveness-performance relationship.

Those findings have clear theoretical implications for the optimal distinctiveness literature. Our study provides further evidence of the importance of exploring heterogeneity in the distinctiveness-performance relationship. Recent studies have especially focused on the role of legitimacy in explaining how this relationship varies, including Taeuscher and Rothe (2021) who tie differences to the existence of alternative sources of legitimacy and Taeuscher *et al.* (2021) who highlight the role of differences in expectations across different audiences. We extend this theoretical debate around the role of legitimacy in determining optimal distinctiveness with our investigation of how legitimation pressures vary depending on product revenue model choices, leading to differences in the distinctiveness-performance relationship across revenue models. Specifically, we explained how legitimacy more quickly erodes with distinctiveness for free products compared to paid products due to greater privacy risk concerns over free products. Products with paid and free revenue models coexist across a variety of industries, and this leads to markedly different implications of distinctiveness for products within the same category. We also discussed how firms can partially attenuate privacy risk concerns and associated legitimacy losses in the case of free products by providing greater transparency about their products' revenue streams, further spotlighting transparency as a legitimacy-enhancing mechanism (e.g., Desai, 2018; Gegenhuber & Dobusch, 2017; Luedicke *et al.*, 2017).

Our findings emphasize how privacy risk, via the potential harm that privacy violations such as data loss may incur for customers, poses unique legitimacy issues in digital markets. In turn, those legitimacy issues point to an emerging class of challenges facing firms in the digital context that a small stream of management research is starting to grapple with (e.g., Casadesus-Masanell & Hervas-Drane, 2015; Kummer & Schulte, 2019). Customer privacy and cybersecurity represents

a complex strategic problem. With ever more avenues for data collection, firms are presented with new opportunities to personalize their products and additional pathways to revenue generation. However, as firms seize those data opportunities, customers' concerns over and exposure to risks in relation to privacy and cybersecurity almost immediately ensue. Naturally, this implies that firms must carefully attend to addressing and managing such issues. Even so, firms ultimately have to strike a balance between the merits of data exploitation and the potential negative impact associated with heightened customer privacy concerns. Where and how firms strike this balance will be consequential for market outcomes in the digital age owing to its effect on legitimacy.

Our work also suggests implications for the business model literature. The choice of revenue model is a critical aspect of business model determination (Massa *et al.*, 2017), reflecting the monetization approach to generate sales from the firm's products (Cassadesus-Masanell & Zhu, 2010). Our work indicates that this choice not only has the obvious implications associated with how revenue will be generated (Rietveld, 2018; Tidhar & Eisenhardt, 2020), but it also has indirect effects on performance via its implications for optimal product positioning. In this respect, our research suggests clear relationships between choices that establish a product's value proposition (positioning) and its revenue model. That is, choices of different revenue models should be accompanied by complementary positioning choices and vice versa. For example, our results indicate that paid products attain higher performance when positioned with a moderate level of distinctiveness while free products attain higher performance with either a high or low level of distinctiveness. We also draw attention to the importance of aligning those revenue models with audience expectations; as part of that effort, we conceptualized *monetization transparency* as an essential characteristic of revenue models. Models with greater transparency (e.g., through privacy statements and freemium) may experience less loss of legitimacy when deviating from category prototypes. In general, our results indicate the importance of closer theoretical integration of the optimal distinctiveness and business model literatures. Indeed, scholars have argued that a central benefit of the business model lens is its emphasis on interdependencies (Lanzolla & Markides, 2021). Future research should examine whether other aspects of business models possess

interdependencies with positioning similar to revenue model choice. If so, these interdependencies will have interesting performance implications that warrant further exploration.

We also add to the platforms literature. Owing to the low barriers to entry and strong network effects (McIntyre & Srinivasan, 2017), digital platforms are increasingly densely contested competitive spaces for firms that develop complementary products. It is not uncommon for categories pre-defined by the platform, in our case divisions in mobile app stores, to harbor thousands of products or more, making it difficult to capture meaningful competitive interactions. For example, while the Apple iOS App Store grew from 500 available apps in 2008 to over 2.2 million available apps in 2017, the number of app store divisions was 20 in 2008 and 24 in 2017. Yet, the literature has still predominantly relied on such pre-defined categories to study intra-platform competition (e.g., Barlow *et al.*, 2019; Eckhardt, 2016; Foerderer *et al.*, 2018). We outline and apply a method based on machine learning for a finer-grained identification of market categories in such contexts. Moreover, we also complement existing work on intra-platform competition. Whereas most attention has been paid to the implications of strategies of platform providers for firms that produce complementary products, such as what happens when platform providers enter into direct competition with complementors (e.g., Foerderer *et al.*, 2018; Wen & Zhu, 2019), we spotlight the important performance implications of fundamental strategic choices (i.e., product positioning and revenue model choice) of those complementor firms.

Our work is not without limitations, which suggest avenues for future research. First, as with many studies, our conclusions are associated with a particular empirical context. But variance in revenue models is not limited to digital platforms, so we encourage future research building on our work to examine the generalizability of our findings to other settings. Within the digital context, we also see potential to consider how the relationships we study might differ outside the context of the Apple App Store. Apple takes an active role in screening apps while other platforms, such as Google Play, do not monitor apps as closely. Relatedly, what might be the implications of whether an app is listed on third-party websites that compile lists of risky apps? These factors may affect the intensity of users' privacy concerns and therefore the degree to which apps pay

legitimacy penalties for deviation from category prototypes. Our specific empirical context also led us to treat app store divisions as non-permeable boundaries when we defined market categories. That is, for tractability of our analyses, we forced all categories to exist within the confines of an app store division. This choice was legitimate in our context given the relatively narrow focus of most apps. However, future research should examine how our results might differ in categories that span multiple app store divisions. It would be interesting to understand how such multi-level categorization schemes affect competition and performance.

We also note that the use of observational data in our empirical analyses means that conclusions of causal relationships in our study are necessarily not absolute. Although the additional analyses we undertook to address endogeneity concerns supported causal conclusions, we encourage additional research that would allow precise identification of the causal effects. We also note that our theoretical arguments related to positioning choices are based on managerial assessments of competitive and legitimation pressures; yet we did not actually observe whether those factors are consciously evaluated by managers when determining product positioning. Future research utilizing fine-grained interview- or survey-based evidence of the cognitions associated with managerial positioning choices would provide valuable insights. Finally, we acknowledge that our work is ultimately interested in economic performance, and although downloads and usage represent critical parts of app performance, additional data, unfortunately unavailable, would have allowed us to home in on the profitability implications of app positioning.

It is also noteworthy to observe that our results suggest that quite some products in digital platforms are positioned non-optimally. This result is somewhat puzzling given the economic incentives to make optimal choices. We encourage future research to investigate explanations for this result. Are developers unaware of the performance implications of their positioning choices, or could they be making choices based on objective functions beyond profit maximization? We also wonder about the role of the particular context of digital platforms here (Yoo *et al.*, 2012). Could a driver of these seemingly non-optimal choices simply be the nascence of some of these markets? If so, we would expect an evolution of positioning choices as digital markets mature.

Alternatively, might it instead have something to do with the fast-paced change in these markets?

Scholars such as Adner *et al.* (2019, p. 253) have suggested that the prevalence of digital markets suggest a need for "a re-examination and expansion of the strategy principles that have guided the field's approach to technological transitions thus far." We suggest that this need may expand beyond just the field's approach to technological transitions. Digital markets feature some unique aspects that may have implications for other fundamental principles in strategic management. As just one example, consider the arguments of Porac *et al.* (1989) about how firms form mental models related to competitive sets or "cognitive communities" of competitive groups. The ability of a Scottish knitwear manufacturer, who faced competition primarily from 17 other local manufacturers, to form a mental model of its competition is likely to be very different from the ability of an app developer to conceptualize its competitive set when entering a category featuring hundreds of competing products or more. What might be the implications for principles of competitive strategy of these new markets featuring low entry barriers, minimal exit barriers, a proliferation of competitors, and privacy concerns?

ACKNOWLEDGEMENTS

We gratefully acknowledge the helpful comments from Associate Editor Gino Cattani, two anonymous reviewers, and conference and seminar participants at the Academy of Management Annual Meeting and the School of Social and Behavioral Sciences of Tilburg University.

REFERENCES

Adner, R., Csaszar, F. A., & Zemsky P. B. (2014). Positioning on a multiattribute landscape. *Management Science, 60*(11), 2794-2815.

Adner, R., Puranam, P., & Zhu, F. (2019). What is different about digital strategy? From quantitative to qualitative change. *Strategy Science, 4*(4), 253-261.

Al-Natour S., Cavusoglu, H., Benbasat I., & Aleem, U. (2020). An empirical investigation of the antecedents and consequences of privacy uncertainty in the context of mobile apps. *Information Systems Research, 31*(4), 1037-1063.

Apple (2020). *Apple services entertain, inform, and connect the world in unprecedented year.*

Barlow, M. A., Verhaal, J. C., & Angus, R. W. (2019). Optimal distinctiveness, strategic categorization, and product market entry on the Google Play app platform. *Strategic Management Journal, 40*(8), 1219-1242.

Blei, D. M., Ng, A. Y., & Jordan, M. J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(1), 993-1022.

Bryce, D. J., Dyer, J. H., & Hatch, N. W. (2011). Competing against free. *Harvard Business Review, 89*(6), 104-111.
Casadesus-Masanell, R. & Hervas-Drane, A. (2015). Competing with privacy. *Management Science, 61*(1), 229-246.
Casadesus-Masanell, R. & Zhu, F. (2010) Strategies to fight ad-sponsored rivals. *Management Science, 56*(9), 1484-1499.
Cennamo, C. & Santalo, J. (2013). Platform competition: Strategic trade-offs in platform markets. *Strategic Management Journal, 34*(11), 1331-1350.
Chen, L., Wang, M., Cui, L., & Li, S. (2021). Experience base, strategy-by-doing and new product performance. *Strategic Management Journal, 42*(7), 1379-1398.
Culnan, M. J. & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science, 10*(1), 104-115.
Deephouse, D. L. (1999). To be different, or to be the same? It's a question (and theory) of strategic balance. *Strategic Management Journal, 20*(2), 147-166.
Desai, V. M. (2018). Collaborative stakeholder engagement: An integration between theories of organizational legitimacy and learning. *Academy of Management Journal, 61*(1), 220-244.
DiMaggio, P. J. & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review, 48*(2), 147-160.
Dunfee, T. W., Smith, N. C., & Ross, W. T. (1999). Social contracts and marketing ethics. *Journal of Marketing, 63*(3), 14-32.
Durand, R. & Paolella, L. (2013). Category stretching: Reorienting research on categories in strategy, entrepreneurship, and organization theory. *Journal of Management Studies, 50*(6), 1100-1123.
Eckhardt, J. T. (2016). Welcome contributor or no price competitor? The competitive interaction of free and priced technologies. *Strategic Management Journal, 37*(4), 742-762.
Farrell, J. & Rabin, M. (1996). Cheap talk. *Journal of Economic perspectives, 10*(3), 103-118.
Foerderer, J., Kude, T., Mithas, S., & Heinzl, A. (2018). Does platform owner's entry crowd out innovation? Evidence from Google Photos. *Information Systems Research, 29*(2), 444-460.
Gegenhuber, T. & Dobusch, L. (2017). Making an impression through openness: How open strategy-making practices change in the evolution of new ventures. *Long Range Planning, 50*(3), 337-354.
Ghose, A. & Han, S. P. (2014). Estimating demand for mobile applications in the digital economy. *Management Science, 60*(6), 1470-1488.
Goldfarb, A. & Tucker, C. E. (2011). Privacy regulation and online advertisement. *Management Science, 57*(1), 57-71.
Haans, R. F. J. (2019). What's the value of being different when everyone is? The effects of distinctiveness on performance in homogeneous versus heterogeneous categories. *Strategic Management Journal, 40*(1), 3-27.
Haans, R. F. J., Pieters, C., & He, Z.-L. (2016). Thinking about U: Theorizing and testing U- and inverted U-shaped relationships in strategy research. *Strategic Management Journal, 37*(7), 1177-1195.
Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal, 29*(2), 147-160.
Hermalin, B. E. & Katz, M. L. (2006). Privacy, property rights and efficiency: The economics of privacy as secrecy. *Quntitative Marketing and Economics, 4*(3), 209-239.
Hintze, M. (2018). Privacy Statements. In E. Selinger, J Polonetsky, and O Tene (Eds.), *The*

Cambridge Handbook of Consumer Privacy (pp. 413-432). Cambridge: Cambriduge University Press.

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

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems, 23*, 856-864.

Hsu, C.-L. & Lin, J.-C. (2015). What drives purchase intention for paid mobile apps? – An expectation confirmation model with perceived value. *Electronic Commerce Research and Applications, 14*(1), 46-57.

Hui, K. L., Teo, H. H., & Lee, S. Y. T. (2007). The value of privacy assurance: An exploratory field experiment. *MIS Quarterly, 31*(1), 19-33.

Kummer, M. & Schulte, P. (2019). When Private Information Settles the Bill: Money and Privacy in Google's Market for Smartphone Applications. *Management Science, 65*(8), 3470-3494.

Lanzolla, G., Markides, C. (2021) A business model view of strategy. *Journal of Management Studies, 58*(2), 540-553.

Liang, K.-Y. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

Lind, J. T. & Mehlum, H. (2010). With or without U? The appropriate test for a U-shaped relationship. *Oxford Bulletin of Economics and Statistics, 72*(1), 109-118.

Luedicke, M. K., Husemann, K. C., Furnari, S., & Ladstaetter, F. (2017). Radically open strategizing: How the premium cola collective takes open strategy to the extreme. *Long Range Planning, 50*(3), 371-384.

Martin, K. D., Borah, A., & Palmatier, R. W. (2017). Data privacy: effects on customer and firm performance. *Journal of Marketing, 81*(1), 36-58.

Massa, L., Tucci, C. L., & Afuah, A. (2017). A critical assessment of business model research. *Academy of Management Annals, 11*(1), 73-104.

McIntyre, D. P. & Srinivasan, A. (2017). Networks, platforms, and strategy: Emerging views and next steps. *Strategic Management Journal, 38*(1), 141-160.

McLachlan, G. & Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering* New York, NY: Marcel Dekker.

Miller, D., Amore, M. D., Le-Breton-Miller, I., Minichilli, A., & Quarato, F. (2018). Strategic distinctiveness in family firms: Firm institutional heterogeneity and configurational multidimensionality. *Journal of Family Business Strategy, 9*(3), 16-26.

Miyazaki, A. D. (2008). Online privacy and the disclosure of cookie use: Effects on consumer trust and anticipated patronage. *Journal of Public Policy and Marketing, 27*(1), 19-33.

Palmeira, M. M. & Srivastava, J. (2013). Free offer ≠ cheap product: A selective accessibility account on the valuation of free offers. *Journal of Consumer Research, 40*(4), 644-656.

Pan, Y. & Zinkhan, G. M. (2006). Exploring the impact of online privacy disclosures on consumer trust. *Journal of Retailing,* 82(4), 331-338.

Pew Research Center (2019). *Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information.*

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

Porac, J. F., Thomas, H., & Baden-Fuller, C. (1989). Competitive groups as cognitive communities: The case of Scottish knitwear manufacturers. *Journal of Management Studies,*

26(4), 397-416.
Porter, M. E. (1996). What is strategy. *Harvard Business Review, 74*(6), 61-74.
Rietveld, J. (2018). Creating and capturing value from freemium business models: A demand-side perspective. *Strategic Entrepreneurship Journal, 12*(2), 171-193.
Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*(4), 573-605.
Schwartz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics, 6*(2), 461-464.
Scott, W. R. (1995). *Institutions and organizations.* Thousand Oaks, CA: Sage.
Shampanier, K., Mazar, N., & Ariely, D. (2007). Zero as a special price: The true value of free products. *Marketing Science, 26*(6), 742-757.
Snihur, Y. & Zott, C. (2019). The Genesis and Metamorphosis of Novelty Imprints: How Business Model Innovation Emerges in Young Ventures. *Academy of Management Journal, 63*(2), 554-583.
Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review, 20*(3), 571-610.
Taeuscher, K., Bouncken, R. B., & Pesch, R. (2021) Gaining legitimacy by being different: optimal distinctiveness in crowdfunding platforms. *Academy of Management J*ournal*, 64*(1), 149-179.
Taeuscher, K. & Rothe, H. (2021). Optimal distinctiveness in platform markets: Leveraging complementors as legitimacy buffers. *Strategic Management Journal, 42*(2), 435-461.
Tidhar, R. & Eisenhardt, K. M. (2020). Get rich or die trying... finding revenue model fit using machine learning and multiple cases. *Strategic Management Journal, 41*(7), 1245-1273.
Tripsas, M. & Gavetti, G. (2000). Capabilities, cognition, and inertia: Evidence from digital imaging. *Strategic Management Journal, 21*(10), 1147-1161.
TUNE (2015). *The State of App Discovery in 2015.*
Wang, Q., Li, B., & Singh, P.V. (2019). Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis. *Information Systems Research,* 29(2), 273-291.
Wang, R. D. & Shaver, J. M. (2015). Competition-driven repositioning. *Strategic Management Journal, 35*(11), 1585-1604.
Wen, W. & Zhu, F. (2019). Threat of platform owner entry and complementor responses: Evidence from the mobile app market. *Strategic Management Journal, 40*(9), 1336-1367.
Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.
Yoo, Y., Boland, R. J., Lyytinen, K., & Majchrak, A. (2012). Organizing for innovation in the digitized world. *Organization Science, 23*(5), 1398-1408.
Zhao, E. Y., Fisher, G., Lounsbury, M., & Miller, D. (2017). Optimal distinctiveness: Broadening the interface between institutional theory and strategic management. *Strategic Management Journal* 38(1), 93-113.
Zhao, E. Y., Ishihara, M., Jennings, P. D., & Lounsbury, M. (2018). Optimal distinctiveness in the video game industry: An exemplar-based model of proto-category evolution. *Organization Science, 29*(4), 588-611.
Zott, C., Amit, R., & Massa, L. (2011). The business model: Recent developments and future research. *Journal of Management, 37*(4), 1019-1042.
Zuckerman, E. W. (1999). The categorical imperative: Securities analysts and the illegitimacy discount. *American Journal of Sociology, 104*(5), 1398-1438.

Figure 1a. Theorized effects of distinctiveness on legitimacy, competition, and performance for paid products

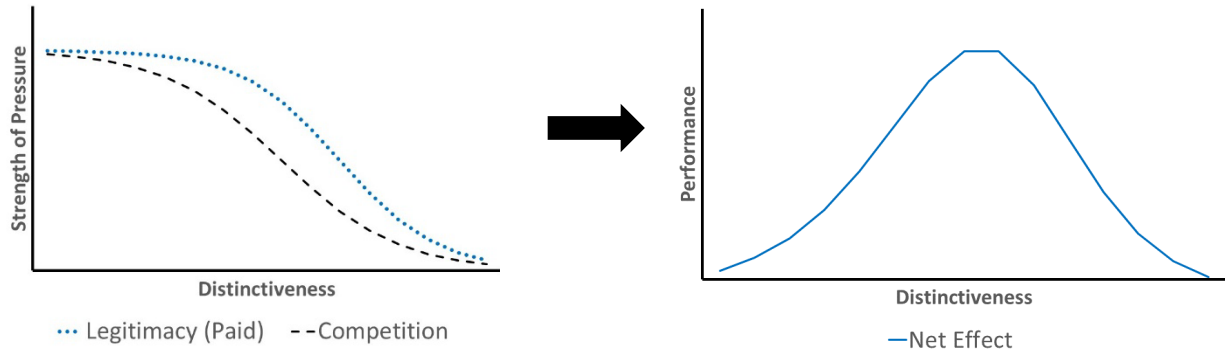


Figure 1b. Theorized shift in the legitimacy curve for paid versus free products

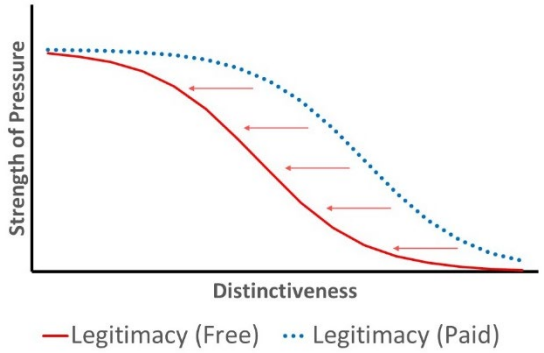


Figure 1c. Theorized effects of distinctiveness on legitimacy, competition, and performance for free products

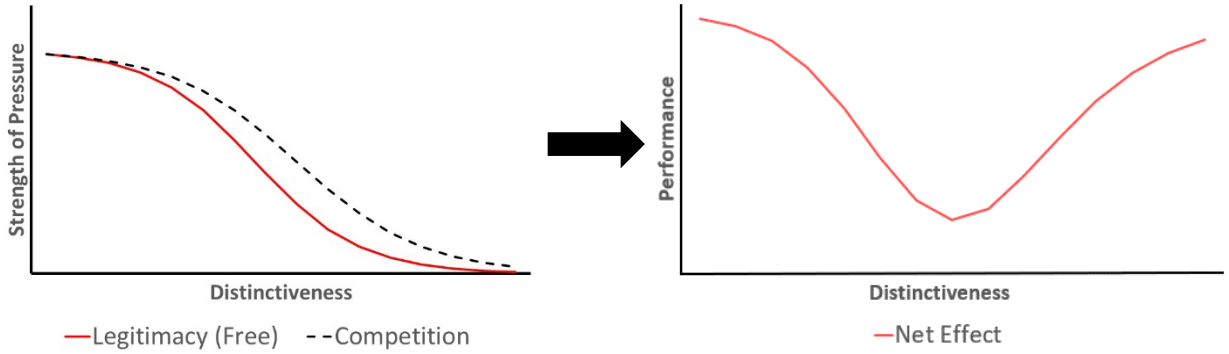
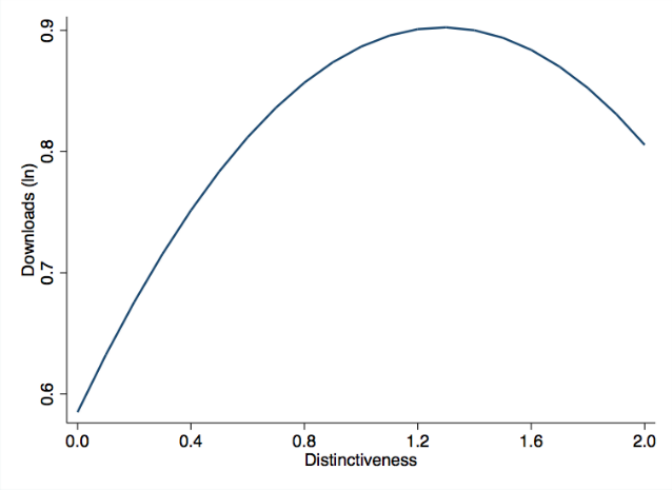


Figure 2. The distinctiveness-performance relationship for paid apps



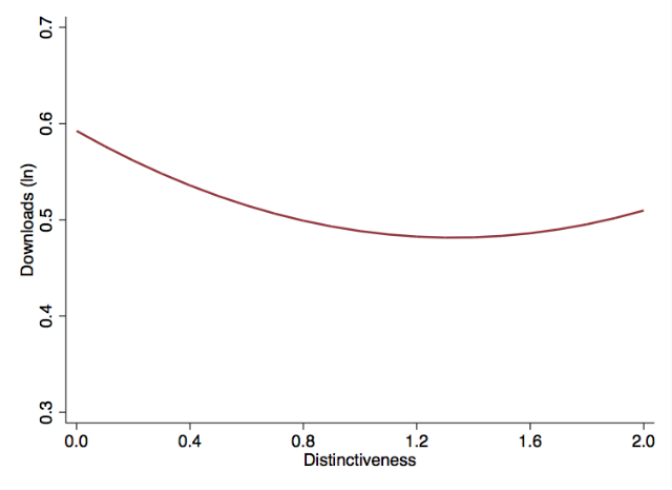


Figure 3b. The distinctiveness-performance relationship for free apps with and without a privacy statement

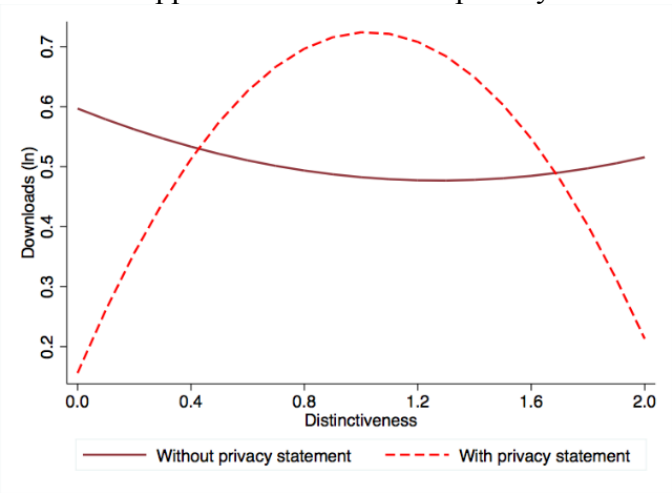


Figure 3c. The distinctiveness-performance relationship for free apps with and without freemium

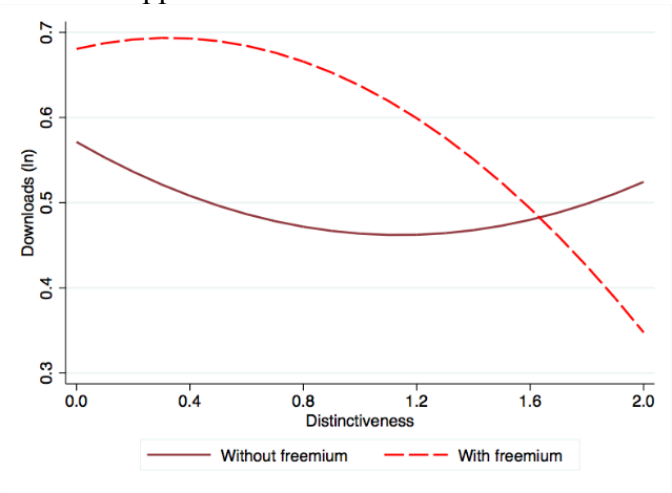


Table 1a. Descriptive statistics and pairwise correlations for paid apps

	Mean	SD	Min	Max	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. Downloads (ln)	0.88	1.71	0.00	12.07																				
2. Distinctiveness	1.01	0.31	0.00	1.92	0.02																			
3. Category heterogeneity	0.11	0.05	0.00	0.50	-0.01	-0.12																		
4. Category maturity	87.28	14.88	1.00	111.00	0.03	0.49	-0.04																	
5. Category size (ln)	4.47	0.71	2.20	7.18	-0.03	0.33	0.13	0.23																
6. Percentage paid in category	0.40	0.20	0.00	1.00	-0.04	-0.47	0.16	-0.29	0.02															
7. HHI	0.47	0.27	0.00	1.00	-0.06	-0.06	-0.00	-0.00	-0.28	-0.08														
8. Developer category experience (ln)	0.53	0.90	0.00	4.33	-0.07	-0.47	0.26	-0.30	0.14	0.46	-0.04													
9. Description length	163.04	131.12	11.00	795.00	0.15	0.02	-0.08	0.11	0.02	-0.04	-0.06	-0.00												
10. Screenshots	3.96	1.29	0.00	6.00	0.04	-0.07	0.03	-0.04	0.05	0.08	-0.01	0.10	0.24											
11. App age	40.12	26.24	0.00	111.00	0.02	0.13	-0.09	0.25	0.02	-0.06	0.01	-0.08	0.15	-0.01										
12. File size (ln)	2.49	1.42	0.00	8.28	0.03	-0.18	0.15	-0.18	0.07	0.19	0.00	0.28	0.08	0.25	-0.27									
13. Price (ln)	0.62	0.79	-0.01	6.91	0.01	-0.03	0.06	-0.03	0.08	0.09	-0.04	0.14	0.20	0.15	-0.07	0.24								
14. Ratings (ln)	1.02	1.75	0.00	12.71	0.45	0.07	-0.06	0.15	0.00	-0.05	-0.03	-0.10	0.27	0.07	0.48	-0.04	0.03							
15. Rating valence between 1 and 2 stars	0.03	0.16	0.00	1.00	0.08	0.02	-0.02	0.03	-0.01	-0.02	-0.00	-0.03	-0.00	-0.04	0.09	-0.06	-0.02	0.14						
16. Rating valence between 2 and 3 stars	0.07	0.25	0.00	1.00	0.08	0.04	-0.04	0.06	-0.00	-0.02	-0.00	-0.04	0.04	-0.05	0.27	-0.10	-0.04	0.34	-0.04					
17. Rating valence between 3 and 4 stars	0.10	0.30	0.00	1.00	0.18	0.06	-0.04	0.09	0.01	-0.03	-0.02	-0.07	0.12	0.03	0.31	-0.06	-0.01	0.47	-0.05	-0.09				
18. Rating valence between 4 and 5 stars	0.11	0.32	0.00	1.00	0.26	0.03	-0.03	0.07	0.01	-0.02	-0.03	-0.05	0.19	0.11	0.15	0.05	0.05	0.50	-0.06	-0.10	-0.12			
19. Recommendations (ln)	0.01	0.05	0.00	3.76	0.08	0.00	-0.00	0.00	-0.00	-0.00	-0.00	-0.01	0.04	0.02	0.02	0.02	0.01	0.10	-0.01	-0.01	0.03	0.09		
20. Ranked	0.02	0.13	0.00	1.00	0.39	0.01	-0.00	0.01	-0.01	-0.02	-0.02	-0.03	0.09	0.03	0.00	0.06	0.04	0.28	0.03	0.01	0.07	0.17	0.09	
21. Featured	0.01	0.05	0.00	1.00	0.12	-0.00	0.01	0.01	0.01	0.01	-0.01	-0.02	0.07	0.03	-0.01	0.04	0.05	0.13	-0.01	-0.01	0.02	0.10	0.08	0.16

Table 1b. Descriptive statistics and pairwise correlations for free apps

	Mean	SD	Min	Max	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Downloads (ln)	0.49	1.92	0.00	15.39																					
2. Distinctiveness	1.06	0.32	0.00	1.92	-0.02																				
3. Privacy statement	0.03	0.15	0.00	1.00	0.09	0.01																			
4. Freemium	0.14	0.35	0.00	1.00	0.18	-0.06	0.24																		
5. Category heterogeneity	0.11	0.05	0.00	0.50	-0.01	-0.16	0.00	-0.02																	
6. Category maturity	85.00	15.10	1.00	111.00	0.01	0.45	0.01	0.04	-0.12																
7. Category size (ln)	4.50	0.78	2.20	7.18	-0.03	0.38	0.01	-0.03	0.13	0.33															
8. Percentage paid in category	0.21	0.17	0.00	1.00	0.08	0.06	-0.01	0.20	-0.15	0.31	-0.06														
9. HHI	0.49	0.31	0.00	1.00	-0.05	-0.08	-0.02	-0.02	-0.07	-0.04	-0.25	-0.04													
10. Developer category experience (ln)	0.42	0.93	0.00	5.78	-0.03	-0.45	0.00	0.02	0.33	-0.27	0.09	-0.17	-0.01												
11. Description length	131.04	110.12	11.00	785.00	0.14	-0.03	0.25	0.25	-0.08	0.06	0.04	0.06	-0.04	0.04											
12. Screenshots	3.79	1.31	0.00	6.00	0.08	-0.00	0.07	0.13	-0.03	0.01	0.02	-0.01	-0.01	0.01	0.25										
13. App age	27.10	22.28	0.00	111.00	0.03	0.07	0.04	0.10	-0.08	0.20	-0.00	0.13	0.01	-0.08	0.09	0.01									
14. File size (ln)	2.74	1.12	0.00	8.30	0.11	-0.02	0.06	0.10	0.04	-0.06	0.02	-0.13	-0.01	0.07	0.15	0.16	-0.26								
15. Ratings (ln)	0.86	1.79	0.00	13.69	0.50	0.00	0.11	0.28	-0.04	0.10	-0.02	0.17	-0.03	-0.05	0.20	0.08	0.46	0.03							
16. Rating valence between 1 and 2 stars	0.01	0.10	0.00	1.00	0.07	0.01	0.00	0.03	-0.01	0.02	-0.00	0.02	-0.00	-0.01	0.01	-0.00	0.06	-0.02	0.11						
17. Rating valence between 2 and 3 stars	0.04	0.20	0.00	1.00	0.10	0.02	0.01	0.05	-0.03	0.07	-0.01	0.08	-0.01	-0.02	0.04	-0.02	0.29	-0.07	0.36	-0.02					
18. Rating valence between 3 and 4 stars	0.07	0.25	0.00	1.00	0.16	0.01	0.05	0.13	-0.03	0.07	-0.01	0.11	-0.01	-0.03	0.09	0.03	0.31	-0.02	0.44	-0.03	-0.06				
19. Rating valence between 4 and 5 stars	0.11	0.32	0.00	1.00	0.29	-0.01	0.07	0.21	-0.02	0.03	-0.02	0.09	-0.02	-0.04	0.14	0.10	0.13	0.08	0.55	-0.04	-0.08	-0.09			
20. Recommendations (ln)	0.01	0.15	0.00	6.96	0.24	-0.01	0.05	0.06	-0.01	0.01	-0.00	0.02	0.00	-0.02	0.08	0.04	0.08	0.06	0.25	-0.01	-0.02	0.09	0.16		
21. Ranked	0.01	0.07	0.00	1.00	0.34	-0.01	0.11	0.08	-0.01	0.00	-0.01	0.00	-0.01	-0.01	0.08	0.04	0.02	0.06	0.24	0.01	0.02	0.05	0.10	0.28	
22. Featured	0.01	0.05	0.00	1.00	0.19	0.00	0.06	0.04	-0.00	0.01	0.00	0.01	-0.00	-0.02	0.07	0.04	0.03	0.05	0.17	0.00	0.01	0.05	0.08	0.17	0.21

Notes. Descriptive statistics and pairwise correlations are based on all observations in the estimation sample (N=829,906 for sample of paid apps; N = 2,410,399 for sample of free apps). Pairwise correlations greater or equal to |0.01| are significant at $p < 0.01$.

Table 2. Generalized estimation equations (GEE) models of the distinctiveness-performance relationship for paid apps

	2.1	**2.2**	**2.3**
Distinctiveness		0.121	0.493
		(0.018)	(0.066)
Distinctiveness^2			-0.191
			(0.037)
Category heterogeneity	0.456	0.465	0.456
	(0.086)	(0.086)	(0.086)
Category maturity	-0.001	-0.001	-0.001
	(0.000)	(0.000)	(0.000)
Category size (ln)	-0.054	-0.070	-0.056
	(0.006)	(0.006)	(0.007)
Percentage paid in category	-0.014	0.040	0.045
	(0.022)	(0.023)	(0.023)
HHI	-0.124	-0.124	-0.125
	(0.007)	(0.007)	(0.007)
Developer category experience (ln)	-0.048	-0.036	-0.030
	(0.005)	(0.006)	(0.006)
Description length	0.001	0.001	0.001
	(0.000)	(0.000)	(0.000)
Screenshots	-0.003	-0.003	-0.004
	(0.003)	(0.003)	(0.003)
App age	-0.015	-0.015	-0.015
	(0.000)	(0.000)	(0.000)
File size (ln)	0.003	0.003	0.003
	(0.003)	(0.003)	(0.003)
Price (ln)	-0.072	-0.074	-0.074
	(0.005)	(0.005)	(0.005)
Ratings (ln)	0.522	0.523	0.523
	(0.008)	(0.008)	(0.008)
Rating valence between 1 and 2 stars	0.100	0.097	0.096
	(0.038)	(0.038)	(0.038)
Rating valence between 2 and 3 stars	-0.304	-0.308	-0.308
	(0.032)	(0.032)	(0.032)
Rating valence between 3 and 4 stars	-0.139	-0.142	-0.143
	(0.032)	(0.032)	(0.032)
Rating valence between 4 and 5 stars	-0.046	-0.049	-0.051
	(0.030)	(0.030)	(0.030)
Recommendations (ln)	0.353	0.352	0.353
	(0.055)	(0.055)	(0.055)
Ranked	1.479	1.478	1.478
	(0.026)	(0.026)	(0.026)
Featured	0.586	0.588	0.586
	(0.074)	(0.074)	(0.074)
App store division dummies	Included	Included	Included
Content rating dummies	Included	Included	Included
Month dummies	Included	Included	Included
Number of observations	829,906	829,906	829,906
Number of paid mobile app	72,017	72,017	72,017
Wald χ^2	38,534	38,621	38,624

Notes. Robust standard errors are in parentheses. The dependent variable is the log-transformed number of monthly app downloads. The constant is estimated but not reported.

Table 3. Generalized estimation equations (GEE) models of the distinctiveness-performance relationship for free apps

	3.1	**3.2**	**3.3**	**3.4**	**3.5**	**3.6**
Distinctiveness		-0.045	-0.167	-0.188	-0.192	-0.204
		(0.012)	(0.041)	(0.041)	(0.040)	(0.040)
Distinctiveness^2			0.063	0.074	0.084	0.090
			(0.022)	(0.022)	(0.021)	(0.021)
Privacy statement	0.191	0.193	0.193	-0.441	0.198	-0.405
	(0.029)	(0.029)	(0.029)	(0.248)	(0.029)	(0.249)
Freemium	0.152	0.151	0.151	0.151	0.109	0.129
	(0.011)	(0.011)	(0.011)	(0.011)	(0.058)	(0.058)
Privacy statement X Distinctiveness				1.295		1.136
				(0.459)		(0.463)
Privacy statement X Distinctiveness^2				-0.613		-0.499
				(0.211)		(0.214)
Freemium X Distinctiveness					0.272	0.229
					(0.127)	(0.129)
Freemium X Distinctiveness^2					-0.207	-0.186
					(0.068)	(0.069)
Category heterogeneity	0.284	0.276	0.271	0.271	0.273	0.273
	(0.047)	(0.047)	(0.047)	(0.047)	(0.047)	(0.0467)
Category maturity	-0.001	0.001	0.001	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Category size (ln)	-0.052	-0.046	-0.051	-0.052	-0.052	-0.053
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Percentage paid in category	0.051	0.046	0.047	0.048	0.043	0.044
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
HHI	-0.027	-0.026	-0.026	-0.026	-0.026	-0.026
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Developer category experience (ln)	-0.052	-0.058	-0.060	-0.060	-0.060	-0.060
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Description length	0.001	0.001	0.001	0.001	0.001	0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Screenshots	0.018	0.018	0.018	0.018	0.018	0.018
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
App age	-0.018	-0.018	-0.018	-0.018	-0.018	-0.018
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
File size (ln)	0.055	0.055	0.055	0.055	0.055	0.055
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Ratings (ln)	0.843	0.842	0.842	0.842	0.842	0.842
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Rating valence between 1 and 2 stars	-0.958	-0.957	-0.957	-0.957	-0.955	-0.955
	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)
Rating valence between 2 and 3 stars	-1.293	-1.292	-1.292	-1.292	-1.290	-1.290
	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)
Rating valence between 3 and 4 stars	-1.213	-1.213	-1.212	-1.212	-1.211	-1.211
	(0.029)	(0.029)	(0.029)	(0.029)	(0.029)	(0.029)
Rating valence between 4 and 5 stars	-1.045	-1.044	-1.044	-1.044	-1.043	-1.043
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Recommendations (ln)	0.151	0.151	0.151	0.151	0.151	0.151
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Ranked	1.627	1.627	1.627	1.626	1.626	1.626
	(0.052)	(0.052)	(0.052)	(0.052)	(0.052)	(0.052)
Featured	0.811	0.811	0.811	0.811	0.812	0.812
	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)
App store division dummies	Included	Included	Included	Included	Included	Included
Content rating dummies	Included	Included	Included	Included	Included	Included
Month dummies	Included	Included	Included	Included	Included	Included
Number of observations	2,410,399	2,410,399	2,410,399	2,410,399	2,410,399	2,410,399
Number of free mobile apps	196,019	196,019	196,019	196,019	196,019	196,019
Wald χ^2	33,641	33,652	33,658	33,668	33,691	33,701

Notes. Robust standard errors are in parentheses. The dependent variable is the log-transformed number of monthly app downloads. The constant is estimated but not reported.