Designing Quality Certificates: Insights from eBay *

Xiang Hui[†]
Washington University

Ginger Zhe Jin [‡]
University of Maryland
and NBER

Meng Liu [§]
Washington University

October 15, 2023

Abstract

Quality certification is a common tool to enhance trust in marketplaces. Should the certification be based on consumer reports, such as ratings, or administrative data on seller behavior, such as the number of seller-initiated cancellations? In theory, incorporating consumer reports makes the quality certificate more relevant for consumer experience but may discourage seller effort because they can be driven by factors not entirely within sellers' control. Alternatively, using administrative data makes the certification more controllable by sellers, but it only tracks a subset of seller behavior and may not be fully aligned with consumer experience. To answer this question, we study a major redesign of eBay's quality certification that removed most consumer reports from its criteria and added administrative data. This change motivates seller effort in dimensions highlighted by the new criteria, but also allows sellers to more precisely target their effort at the threshold. Buyers place a higher value on the quality certificate and are more likely to purchase again on the platform in markets where administrative data is more correlated with consumer reports. Lastly, the proportion of certified sellers becomes more homogenized across markets and sales become more concentrated towards large sellers.

Keywords: quality certificate, reputation, moral hazard, platform, e-commerce.

*We are grateful to eBay for providing access to the data. We thank Andrey Fradkin, Michael Luca, and seminar and conference participants at Boston University, the NBER Summer Institute, BU Platform Strategy Symposium, TSE seminar on the Economics of Platforms, Washington University Marketing Brownbag, and Amazon economics seminars for constructive feedback. None of us has a financial relationship with eBay. The content and analyses in this paper reflect the authors' own work and do not relate to any institution or organization with whom the authors are affiliated. All errors are our own.

[†]hui@wustl.edu
[‡]ginger@umd.edu
[§]mengl@wustl.edu

1 Introduction

Trust precedes economic prosperity. To build trust in markets with asymmetric information on product quality, market designers typically use information mechanisms. The most common information mechanism is seller reputation for good behavior in past sales (Shapiro, 1983), which has been shown to increase market efficiency in many contexts (Tadelis, 2016). Another popular but less studied mechanism is quality certification (Dranove and Jin, 2010). The essence of quality certification is that some trusted third party endorses firms with a certificate if their underlying quality is above some predetermined threshold, allowing consumers to make more informed decisions (Leland, 1979). Real-world examples of quality certificates include the Better Business Bureau ratings of businesses and charities, U.S. News ranking of colleges, and various seller badges on e-commerce platforms (e.g., Airbnb's Superhost badge).

How to measure quality is a key question in designing quality certificates. Typically, the measurement is based on two types of information: consumer reports and administrative data. Consumer reports, such as ratings and reviews, are user-generated content that summarizes consumers' opinions about their transaction experience overall or in specific dimensions. Administrative data, such as the number of seller-initiated cancellations and handling time based on tracking information provided by post offices, is automatically collected for record keeping by market regulators (e.g., an e-commerce platform), and focuses more on seller behavior and less on consumer opinions. Theoretically, there is a relevance–controllability trade-off between using these two types of information: While using consumer reports makes the certification more relevant for consumer experience, they may arise from factors not entirely within sellers' control, which could discourage seller effort. Administrative data, on the other hand, increases sellers' controllability in managing the certification, but it tracks only a subset of seller behavior and therefore may not be fully aligned with consumer experience. What are the consequences of using consumer reports vs. administrative data in constructing quality certificates on seller behavior, consumer satisfaction, and market outcomes?

Given the theoretical ambiguity, we empirically study these questions by leveraging rich data from eBay. We choose eBay as our empirical context for two main reasons. First, in terms of identification, in 2016, eBay removed most consumer reports and added administrative data in the criteria for the eBay Top Rated Seller (eTRS) badge. This major re-design creates a rare exogenous shock to the type of information used in the quality certificate in a large e-commerce marketplace.

Second, in terms of analyzing the underlying mechanisms, eBay has a large number of product markets that are differentially affected by the one-size-fits-all regime change because they differ in the relevance (of the administrative data for buyers) and controllability (of consumer reports by sellers) trade-off due to the nature of the different products. This cross-market variation sheds light on the key trade-off between using consumer reports and administrative data in designing quality certification in different types of markets.

We start by constructing proxies for relevance and controllability. Conceptually, sellers' controllability (in managing their certification status) concerns how likely they will get certified conditional on good behavior. The uncertainty arises in the old regime because certification depends on consumer reports, which is only partially determined by seller behavior.[1] We proxy for controllability by the probability that a seller receives a positive consumer report conditional on meeting the requirements of the new certification regime (which focuses on seller behavior).[2] We then compute the controllability measure for each product market following product subcategories defined by eBay. Essentially, a low-controllability market is one where sellers are likely to receive negative consumer reports despite good seller behavior, and therefore less able to manage their certification status under the old regime.

Next, relevance conceptually captures how aligned the new certification is with consumer experience. Towards a proxy, we use the correlation coefficient between a seller's actual certification status under the old regime and their hypothetical certification status under the new regime at the policy announcement (i.e., apply the new requirements on the seller's past behavior). Intuitively, a high-relevance market has both regimes giving certification to a similar set of sellers, which means that the new certification aligns well with consumer reports (reflected in the old certification) in that market.

We have three key findings. On the seller side, increasing sellers' controllability in quality certification motivates their effort in dimensions highlighted by the new requirements. However, this change allows sellers to more precisely target their effort at the certification threshold. On the buyer side, in markets where the administrative data are more aligned with consumer reports (i.e., high-relevance markets), consumers are more likely to buy items with the quality certificate and are more likely to purchase again on the platform six months later. Lastly, at the market level, the

[1] An example is logistic delay due to weather: consumers may leave negative ratings because they did not receive their items on time, even though the delay is caused by unexpected weather condition despite sellers' on-time handling.

[2] We define "good" seller behavior based on eBay's new certification requirements because we think that eBay has a good idea about what consumers value the most based on many A/B tests. While there are different ways of defining good behavior, our proxy is valid if these definitions are positively correlated, which we believe is the case.

proportion of certified sellers is more homogenized across markets and sales are more concentrated towards larger sellers especially in low-controllability markets under the old regime.

Our results yield a few insights for market designers. First, a good design of quality certification depends on a thorough evaluation of the trade-off between using administrative data to achieve high seller controllability and using consumer reports to achieve high relevance to consumer satisfaction. Second, the net effect of increasing seller controllability on quality provision is ambiguous due to the threshold effect. Third, whether consumers find the quality certification useful depends on its alignment with consumer reports. Lastly, the design of quality certification can have long-run effects on market concentration. We elaborate on these implications in the conclusion.

1.1 Literature Review

Our paper contributes to two strands of empirical literature. The first strand studies the design of quality disclosure and certification. Jin and Leslie (2003) show that mandatory display of hygiene quality grade cards motivates restaurants to improve hygiene quality. Since the paper, more papers use natural experiments to study the effect of information disclosure on seller behavior (see summaries of these papers in Dellarocas (2003); Dranove and Jin (2010); Einav et al. (2016)). In the eBay context, how much consumers value a seller's eTRS badge is equivalent to a 7% increase in their willingness to pay on the U.K. site (Elfenbein et al., 2015) and a 3% increase in sales price on the U.S. site (Hui et al., 2016). In terms of designing the eTRS certificate, research shows that both the stringency and granularity of the certification threshold can have large impacts on seller entry and quality provision on eBay (Hui et al., 2017, 2020). Outside eBay, Farronato et al. (2020) show that occupational licensing adds little information value above and beyond consumer ratings on a platform that offers home services. In comparison, Jin et al. (2022) find that partial and mandatory licensing of food sellers on Alibaba, following the 2015 Food Safety Law of China, improved the average quality of surviving food sellers. Our paper contributes to this literature by studying another key design question in quality certification: What are the benefits and drawbacks of using consumer reports vs. administrative in constructing quality certification?

Our paper is also related to a larger literature that studies consumer reports and reputation systems. The literature finds that reputation systems lead to higher demand in various contexts.[3]

[3]The contexts include e-commerce (Chevalier and Mayzlin, 2006; Vana and Lambrecht, 2021; Park et al., 2021), online labor markets (Pallais, 2014; Barach et al., 2020; Benson et al., 2020), review websites (Luca, 2016), hotel websites (Hollenbeck et al. (2019)), video games (Zhu and Zhang (2010)), consumers' choice of food (Bollinger et al. (2011); Bai (2018)) and eBay (Dewan and Hsu, 2004; Resnick et al., 2006; Cabral and Hortacsu, 2010; Saeedi, 2019). Ratings improve consumer welfare (Wu et al., 2015; Lewis and Zervas, 2016; Reimers and Waldfogel, 2021).

However, a more recent literature shows that consumer reports may be imperfect, as buyers may fear retaliation from sellers (Dellarocas and Wood, 2008; Bolton et al., 2013; Fradkin et al., 2019), some reviews are fake (Anderson and Simester, 2014; Mayzlin et al., 2014; Luca and Zervas, 2016; He et al., 2020), and ratings may be influenced by sales price, inflated, or left selectively (Luca and Reshef, 2021; Filippas et al., 2022; Anderson and Sullivan, 1993; Dellarocas et al., 2006; Hu et al., 2009; Moe and Trusov, 2011; Nosko and Tadelis, 2015; Ho et al., 2017; Ishihara and Liu, 2017; Brandes et al., 2019; Hui et al., 2022). The proposed solutions include moving from bilateral to unilateral rating systems (Klein et al., 2016; Hui et al., 2018), changing the timing of review revelation Fradkin et al. (2021), and creating incentives for reviews (Cabral and Li, 2015; Li et al., 2022; Burtch et al., 2018; Marinescu et al., 2021; Karaman, 2021; Fradkin and Holtz, 2021). Our paper contributes to this literature by showing that using consumer reports could discourage seller effort. Therefore, when constructing trust signals, market designers need to weight the informational benefit of consumer reports against its negative effect on effort incentives. These results shed light on how to better aggregate different types of information, such as the first steps taken by Dai et al. (2018) and Vatter (2021).

2 Background and Conceptual Framework

2.1 Background

The quality certification program on eBay is called eBay Top Rated Seller (eTRS). On the 20th of each month, all sellers are evaluated against a set of requirements that are publicly known to market participants. At the high level, eBay aggregates a number of seller performance metrics to determine whether the seller meets the threshold for the eTRS certificate. We will discuss the breakdown of the metrics in Table 1.

The main benefit of having an eTRS certificate is that sellers can have the eTRS badge displayed on their listings on the search results page (e.g., Figure 1) and on the listing page (shown after clicking on the item on the search results page). This benefit is valuable because the eTRS certificate is the only reputation signal that is shown on the search results page, which can affect consumers' search and purchase decisions. Besides the eTRS badge, consumers can also observe other reputation metrics with a few clicks: clicking on an item on the search result page (e.g., Figure 1) leads consumers to the listing page (e.g., Figure 2), where consumers can see a seller's percentage

See Dellarocas (2003) and Tadelis (2016) for summaries.

Figure 1: Example: Search Results Page

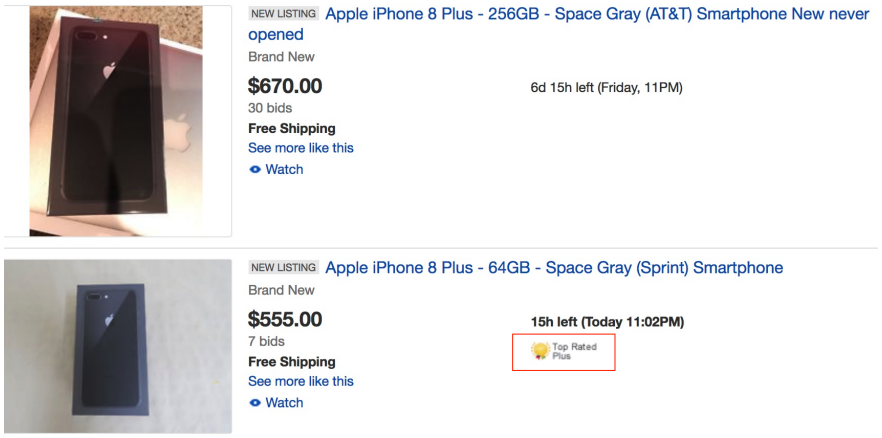
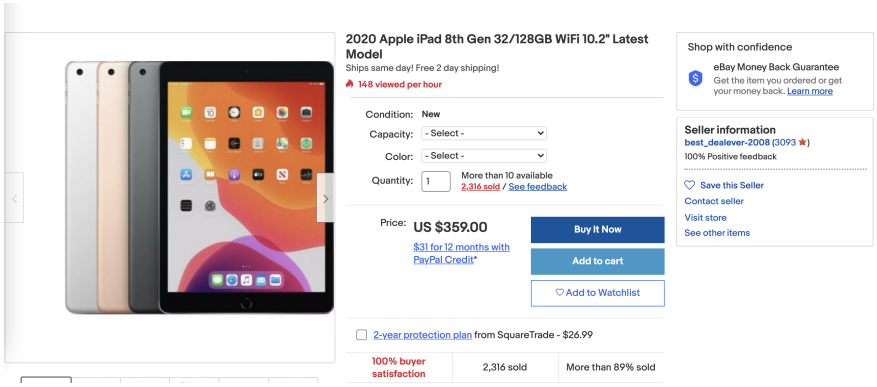
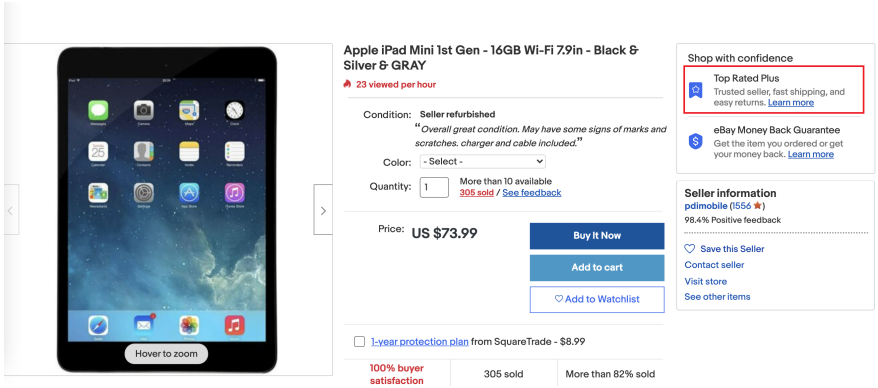


Figure 2: Example Listing Pages



(a) Example 1: A listing without an eTRS plus badge



(b) Example 2: A listing with an eTRS plus badge

positive ratings. On the listing page, an additional click on the seller profile name leads consumers to the seller profile page, where consumers can see a seller's Detailed Seller Ratings (DSRs) in four dimensions: item as described, communication, shipping speed, and shipping charges.

Note that the eTRS badge is the most salient and important quality signal on eBay. In particular, Nosko and Tadelis (2015) show that the distribution of sellers' percentage positive ratings is very skewed and has a median of 100% and a mean of 99.4%, which limits its effect on established sellers as used in this study. Additionally, using click stream data from eBay, Nosko and Tadelis (2015) show that less than 1% of users ever go to the seller's profile page, which implies that DSRs are not very visible to most users.

Table 1: Changes in eBay's eTRS Certification Requirements

Before	After
Certification Requirements:	*Certification Requirements*:
1) Negative or neutral feedback (C)	
2) Low Detailed Seller Rating on item as described (C)	
3) Buyer claims (C)	3') Unresolved buyer claims (A)
4) Seller cancellation (A)	4') Seller cancellation (A)
5) Low Detailed Seller Rating on shipping time (C)	5') Late delivery based on tracking information (A)
Threshold:	*Threshold*:
- defect (metrics 1–5) rate $\leq 2\%$	- defect (metrics 3', 4') rate $\leq 0.5\%$
- late delivery (metric 5) subsumed in defect	- late delivery (metric 5') rate $\leq 5\%$

Notes. Here (A) and (C) denote administrative data and consumer reports, respectively. Low Detailed Seller Ratings are scores of 1 or 2 on a five-point scale. The threshold on the new defect rate and late delivery rate were set at a level so that eBay expected no significant change in the number of eTRS sellers after the policy change. The requirement on minimum past sales is unchanged: $1,000 in value and 100 transactions in the past 12 months.

In 2015, eBay re-designed the eTRS program, aiming to create a simpler and more objective standard. The new eTRS criteria was announced in September 2015 and implemented in February 2016, and the change is summarized in Table 1.[4] At the high level, consumers reports (metrics 1, 2, 3, 5) were replaced with administrative data that emphasizes seller behavior. More specifically, under the new regime, a buyer claim counts as a defect only if the sellers are unable to resolve them (change from metric 3 to 3')[5]; and late delivery is now based on tracking information provided by post offices instead of consumer reports (metric 5' instead of 5). As a result of the regime change, sellers should find it easier to manage and predict their future eTRS status. For example, a seller eager to earn the eTRS badge may offer refunds in case of claims (which would be recorded as

[4]See the cached (historical) announcement page on 09/11/2015 at http://bit.ly/3t93hPm.

[5]When a buyer files a claim on an order and the seller cannot resolve it in three days, the case is escalated to eBay. It is counted as an unresolved buyer claim if eBay decides the seller is at fault. While buyer claims are still consumer reports, the new regime excludes resolved buyer claims and thus puts more emphasis on seller effort to resolve buyer claims.

resolved claims by eBay), refrain from cancellation, and send the items to post offices within the specified handling time.

Note that even after the policy change, percentage positive ratings (metric 1) and DSRs (metrics 2 and 5) of a seller are still visible to buyers on the listing page and seller profile page. As previously discussed, however, the eTRS certification is much more salient and important than other reputation signals for established sellers in our sample, which justifies our focus on the eTRS certification.

2.2 Conceptual Framework

What would be the impact of using consumer reports vs. administrative data in quality certification? We argue that the theoretical trade-off is relevance vs. controllability. On the one hand, consumer reports are relevant for consumers in that they reflect their experience in transactions. However, due to reasons discussed in Section 1.1, consumer reports can be noisy from the sellers' perspective because their effort can only partially determine consumer reports. The lack of controllability of their eTRS status under the old regime (and hence payoffs) discourages seller effort (Baker, 1992). On the other hand, using administrative data on seller behavior increases sellers' controllability of their eTRS status but may not fully align with consumers' transaction experience, thereby possibly reducing consumer valuation of the certificate.

The policy change replaces consumer reports with administrative data in quality certificates. What would happen in the marketplace? On the buyers' side, if seller effort does not change, buyers may value the new quality certificate less than the consumer-reports-based quality certificate, and the valuation should decrease more in markets where the administrative data are less relevant (i.e., less aligned with their transaction experience). On the sellers' side, various forces shape their behavior. First, sellers can better predict how their effort translates into their certification status, and hence their economic payoffs. Therefore, they are motivated to exert effort according to the classical principal-agent theory. Second, the discrete nature of quality certification may affect sellers' effort decisions depending on sellers' distance to the threshold. Third, seller effort can be multidimensional, so the improvement in some effort dimensions may come at the cost of other dimensions (Holmstrom and Milgrom, 1991).

In equilibrium, if consumers incorporate changes in seller effort in response to the policy change, they may put a higher value on the new quality certificate even if it is less focused on measures of consumer experiences. In that world, the policy change could lead to higher revenue as sellers exert more effort, consumers receive better services, and both sides are more willing to trade on

eBay. The converse would be true if the change reduced the information value of the certification for consumers and sellers were motivated to improve on less consumer-relevant dimensions. In an extreme, if consumers find the new quality certificate to be irrelevant and hence not valuable, sellers may decrease their effort and abandon the certificate altogether. Either way, the policy change could lead to a different composition of sellers and market concentration, depending on how different seller types benefit from the certification. Because of the ambiguity of theoretical results, we empirically study the eTRS re-design to shed light on the research question.

3 Data

We use proprietary data from eBay, which is a popular e-commerce marketplace.[6] Our starting point is the set of all listings on eBay from 11 months before to 11 months after the month when eBay announced the eTRS regime change, excluding the listings in Motors and Real Estate categories. We define three periods based on the policy announcement date and implementation date: "before" refers to the 11 months before the policy announcement (October 20, 2014 to September 19, 2015); "interim" refers to the 5 months between the policy announcement and its implementation (September 20, 2015 to February 19, 2016); "after" refers to the 6 months after the policy implementation (February 20 to August 19, 2016).[7] For convenience of reference, Month 0 denotes the first month of the new eTRS policy implementation (February 2016), and all other months are normalized accordingly. For example, the policy announcement month is Month -5, and the sixth month after the policy implementation is Month 5.

To focus on professional sellers, we study those who had sold at least $5,000 in the year before the policy announcement. We then keep sellers who listed at least one item in each of the before, interim, and post periods, to mitigate the potential problem of dynamic entry and exit.

While we cannot report summary statistics of the raw data to keep eBay's business data confidential, we report normalized time series of key variables where the normalization is taken with respect to the corresponding values in the first month in our sample. More specifically, we plot the time series by the four types of markets, which we will define below, in Appendix F.

Because our goal is to understand the trade-off of using different metrics, we first study the correlation between the various metrics based on consumers reports and administrative data. We

[6] eBay was the second most visited e-commerce marketplace in 2021 (https://www.webretailer.com/b/online-marketplaces/ (accessed August 2022).

[7] The policy announcement date is September 11, 2015. Therefore, the first month in the interim period is the first full month after the policy announcement.

create a heat map of these metrics in Figure 3, where a darker color indicates a higher correlation coefficient between the two metrics. The correlations are calculated at the transaction level based on data from the three months before the policy announcement. All quality measures are dummy variables. For example, Low DSR on Item Description equals 1 if the transaction received a score of 1 or 2 on a five-point scale on that DSR. Similarly, Low DSR on Shipping Speed is 1 if the transaction receives a low Detailed Seller Rating on shipping speed from consumers.

Figure 3: Correlation Coefficients between Consumer Reports and Administrative Data

	(A)	(B)	(C)	(D)	(E)	(F)	(G)
(A) Negative Feedback	1.00	0.55	0.18	0.38	0.04	0.16	0.01
(B) Low DSR on Item Description		1.00	0.12	0.34	0.02	0.12	0.00
(C) Buyer Claim			1.00	0.14	-0.01	0.58	0.04
(D) Low DSR on Shipping Time				1.00	0.03	0.10	0.05
(E) Seller Cancellation					1.00	0.00	-0.03
(F) Unresolved Buyer Claim						1.00	-0.01
(G) Late Delivery (Tracking Info)							1.00

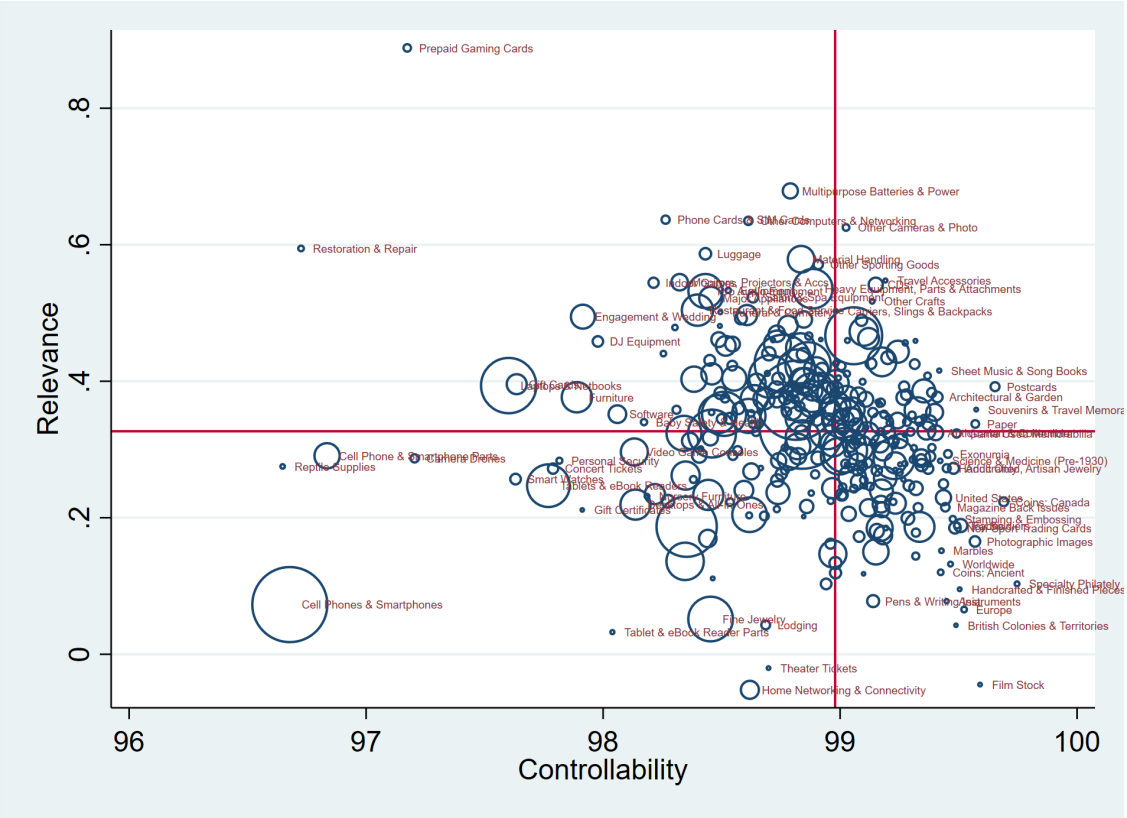
Notes: Correlations are calculated at the transaction level based on data from the three months before the policy announcement. All quality measures are dummy variables.

The figure suggests a few patterns: first, the consumer reports metrics, namely negative feedback, low DSR on item description, buyer claim, and low DSR on shipping speed, are positively correlated with each other. For example, negative feedback is highly correlated with low DSR on item description, suggesting that a major source of customer complaint is from unclear item description. Second, the correlations between consumer reports and administrative data on seller behavior (i.e., seller cancellation, unresolved buyer claims, and late delivery based on tracking information provided by post offices) are weak in general, suggesting that consumer reports cannot be fully substituted by administrative data and therefore the relevance of administrative data to consumer experience is an important margin in our analysis. An exception here is the positive correlation between buyer claim and unresolved buyer claim, which comes from the fact that the latter is a subset of the former. Third, the three dimensions of the administrative data are barely correlated, suggesting that they measure different aspects of seller quality.

Next, to understand the trade-off of using consumer reports and administrative data, we categorize a large number of product markets[8] into different groups depending on the relevance and controllability measures. To remove outliers, we remove all markets with less than $300,000 in sales in the first three months of the sample period. This procedure leaves us with 328 markets, which

[8]The product markets follow product subcategories defined by eBay. Examples of subcategories include Video Games, Women's Clothing, and Cell Phones & Smartphones.

Figure 4: Markets Differ in Controllability and Relevance



Notes: Controllability is the probability of receiving a positive consumer report conditional on having good seller behavior. Relevance is the correlation coefficient between a seller's actual and simulated badge statuses on the day before the policy announcement date. Simulation is done by applying the new requirements on seller performance before the policy announcement. Circle size represents total dollar sales.

accounts for 99.7% of the total sales.

We present the categorization of markets by controllability and relevance in Figure 4. Each circle represents a product market, with circle size representing total dollar sales. To construct the proxy for controllability, we use the probability of receiving a positive consumer report (i.e., items 1), 2), 3), and 5) in the old certification requirements) conditional on having good seller behavior based on the administrative data (i.e., all items in the new certification requirements) in that market. Note that both dollar sales and conditional probability are calculated using data from the three months before the policy announcement date. The horizontal axis values are in percentage points. Intuitively, a low controllability in a market means that sellers are likely to receive negative consumer reports despite good seller behavior, and this lowers their controllability in managing their certification status under the old regime in that market.

To construct the proxy for relevance, we use the Pearson correlation coefficient between a seller's

certification status under the old regime and what would have the status been had the policy change happened immediately at the its announcement. The simulated badge status is done by applying the new requirements to seller performance immediately at the policy announcement, so that the sellers haven't had the chance to react to the new regime. Note that the correlation coefficients are also calculated using data from the three months before the policy announcement date. Intuitively, a high relevance in a market means that both regimes give certification badges to similar sellers, which in turn means that buyers should find the new certification to align well with consumer reports as required in the old certification in that market.

Based on the median split of the two measures (the two lines in Figure 4), we group markets into four types. For example, the Prepaid Gaming Cards category ranks high on relevance, and the Cell Phones & Smartphones category ranks low on controllability. To understand the characteristics of markets that differ in controllability and relevance, we define a dummy variable for whether the relevance measure of a market is above median, and another dummy for above-median controllability. We regress the two dummies on the following market-level characteristics: logged average sales price, whether the share of standardized products (inferred from the existence of an internal eBay Product ID) in a market is above median, logged average shipping package size in cubic inches, logged Herfindahl–Hirschman Index (or, HHI, which ranges from 0 to 10,000), and market size in terms of logged quantity sold.

The results are shown in Table 2. Column 1 shows that relevance is positively correlated with having more standardized products in the product category. This correlation is intuitive: in markets with less standardized products, ratings are more likely to capture consumers' idiosyncratic taste, which is harder to measure from administrative data on seller behavior. The correlation can explain why markets of contain highly standardized products (e.g., Prepaid Gaming Cards) score high on relevance. Column (2) shows that larger package size is also positively correlated with relevance. This is because a higher shipping cost likely increases the correlation between delay in seller's handling time and low consumer ratings on shipping. The two findings are robust to the inclusion of HHI and market size in the regression. Next, columns (4)–(6) show that seller controllability in a market is negatively correlated with price. This correlation can explain why markets such as Cell Phones & Smartphones have low controllability. Specifically, this result is consistent with the previous literature that found a higher sales price leads to higher consumer expectation and, therefore, more critical ratings conditional on seller behavior (e.g., Luca and Reshef (2021)). The fact that these regression results are intuitive increases our confidence on market categorization.

Table 2: Predictors of Controllability and Relevance

	(1)	(2)	(3)	(4)	(5)	(6)
	relevance	relevance	relevance	control	control	control
log(price)	-0.002	-0.009	-0.011	-0.242***	-0.250***	-0.239***
	(0.008)	(0.008)	(0.009)	(0.027)	(0.029)	(0.030)
standardized	0.040***	0.039***	0.038**	-0.029	-0.030	-0.031
	(0.015)	(0.015)	(0.016)	(0.054)	(0.054)	(0.056)
log(package size)		0.014*	0.020***		0.019	-0.015
		(0.007)	(0.007)		(0.026)	(0.026)
log(HHI)			0.020***			-0.116***
			(0.007)			(0.023)
log(total quantity)			0.012**			-0.055***
			(0.006)			(0.020)
Constant	0.307***	0.254***	-0.008	99.807***	99.735***	101.068***
	(0.031)	(0.041)	(0.109)	(0.111)	(0.150)	(0.387)
Observations	325	325	325	326	326	326
R-squared	0.022	0.033	0.062	0.205	0.206	0.263

Notes: One observation is a product category. Outcome variables are dummies for whether the relevance and controllability of a market are above median. Standardized is a dummy for whether the share of standardized products in a product category is above median. Package size is measured in cubic inches. HHI is in total dollar sales. *** p<0.01, ** p<0.05, * p<0.1.

4 Results on Sellers

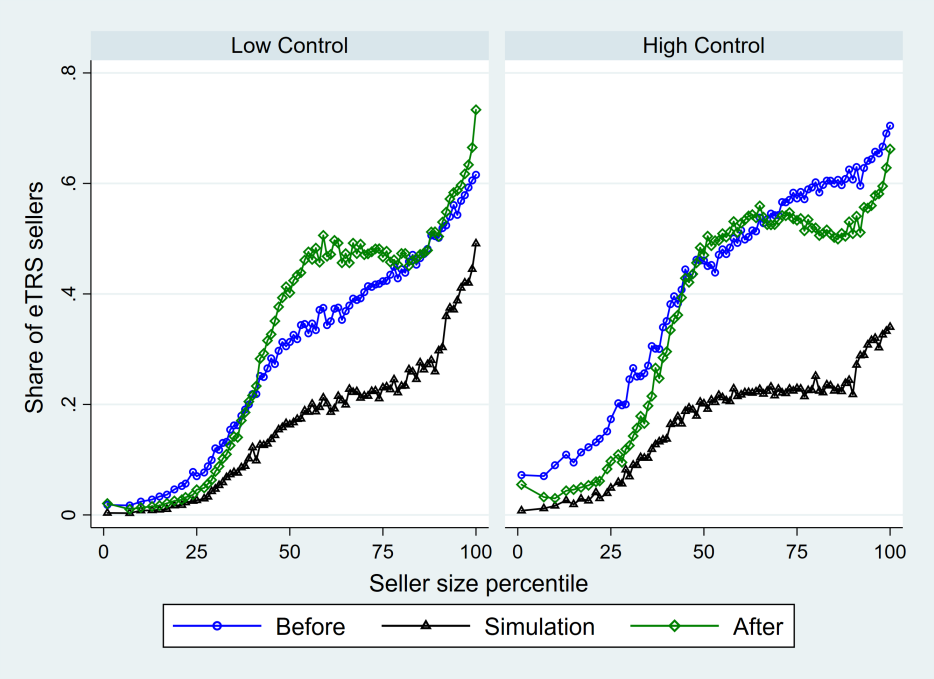
4.1 Selection Effect and Changes in Behavior

The overall effects of the regime change on sellers can be decomposed into (1) an immediate selection effect, i.e., changes in the composition of certified sellers before any changes in seller behavior and (2) seller behavior changes in response to the new policy. To measure the selection effect, we simulate a seller's hypothetical eTRS status by applying the new requirements to the seller's performance metrics on the policy announcement date (September 2015, or Month -5). At that time, sellers had not yet had an opportunity to change their efforts in response to the new policy, so the difference between the seller's actual and simulated eTRS status should capture only selection.[9] To measure the changes in seller behavior in response to the new policy, we focus on the difference between a

[9]Note that there could be measurement errors when we simulate a seller's eTRS status. For example, we observe the number of unresolved claims (or other certification metrics) that a seller received. If eBay first thought the seller is at fault but reversed the decision after the seller appealed with more evidence, that transaction should no longer be counted as a defect by eBay. However, we would still label this instance as a defect because we do not observe metrics revision in our sample. In the presence of measurement errors, while we can not be certain of the absolute magnitude of the selection effect, we can still study how the relative selection effect differs by market type assuming the measurement error is independent of the above two metrics. Additionally, we can still study the changes in seller behavior over time if we assume that the measurement error is constant over time.

seller's simulated eTRS status on the policy announcement date and the actual eTRS status on the policy implementation date. For example, if a seller is not eTRS certified by simulation, but later becomes eTRS certified, we infer that the seller has exerted effort based on the eTRS requirements after the policy announcement.

Figure 5: Policy Effect on Sellers by Controllability



Notes: The two markets are defined based on a median split on the seller controllability measure. Simulated share of eTRS sellers (represented by triangles) is normalized by a constant. "Before" refers to Month -5; simulation is done at Month -5. "After" refers to Month 5.

In Figure 5, we plot the selection effect and changes in seller behavior by seller controllability in a market. For clarity, we present seller-side results by markets with different controllability, because controllability is more applicable for sellers' effort decisions; we also present the same analyses by markets of different controllability-by-relevance in Appendix B and the insights are similar. In the figure, we plot the "before" (Month -5), simulated (at Month -5),[10] and "after" (Month 5) shares of eTRS sellers across different percentiles of seller sizes, where seller size is measured by a seller's historical sales by the policy announcement date. We control for seller size on the x-axis because (1) randomness in consumer reports differentially affects sellers with different sizes, per the law of large numbers, and (2) the eTRS badge is applicable only to sellers exceeding a minimum number of past sales. Therefore, sellers of different sizes would expect different benefits from the eTRS

[10] To comply with the data agreement, we shift all curves by a constant (i.e., the same constant for both sub-figures).

badge. We elaborate on why seller size matters in Appendix A.

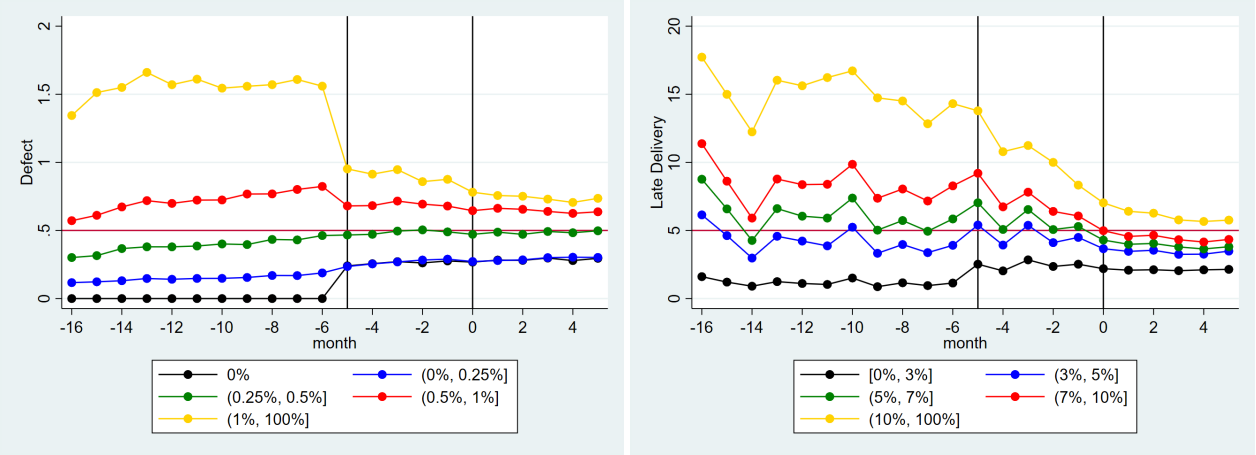
Figure 5 shows that the selection effect is generally negative, because the simulated eTRS share is smaller than the actual eTRS share under the old regime for all seller sizes and both market types. This is expected because sellers have optimized their behavior based on the old certification requirements but have not yet readjusted their behavior towards the new requirements right at the policy announcement. More importantly, the selection effect is more positive (or less negative) in low-controllability markets than in high-controllability markets, as the negative gap between the simulated curve and the "before" curve for all seller sizes is smaller in low-controllability markets. We provide a detailed discussion of market type, seller size, and the selection effect in Appendix B. As can be seen from the graph, the smaller negative gap is jointly explained by a lower "before" curve and a higher simulated curve in low-controllability markets: The reason for a lower "before" curve is that, fixing effort, sellers were less likely to be badged in low-controllability markets before the policy change. The reason for a higher simulated curve is that in low-controllability markets, sellers that intend to be badged need to overshoot on administrative data that they can control to overcome the randomness in consumer reports. We present more evidence on this overshooting in Appendix C).

Next, we infer changes in seller behavior from the difference between the "after" and simulated shares of certified sellers. From the figure, we see that sellers overall improve their effort based on the new certification requirements. However, despite the more positive selection effect for sellers in low-controllability markets, we do not see a larger share of eTRS sellers after the policy implementation in these markets; graphically, this is shown as a higher simulation curve but a similar "after" curve across the two types of markets. This observation suggests that the quality improvement is smaller among sellers that operate in low-controllability markets. This result can be surprising at first glance, because the classical principal-agent theory would predict a greater incentive to exert effort when effort is more observable to the principal (Baker, 1992). However, among other forces discussed in Section 2.2, this argument ignores the binary nature of certification: once a seller has reached the minimum quality threshold in certification, there is no benefit of exerting additional effort. This behavior gives rise to a threshold effect, which is the phenomenon that sellers target their effort level so that they just pass the threshold— a behavior that has also been documented in other contexts, e.g., (Hunter, 2020; Alé-Chilet and Moshary, 2022). Given that the two "after" curves look similar, we hypothesize that the smaller quality improvement in low-controllability markets is a result of the threshold effect, which we test in the next subsection.

4.2 Threshold Effect

To test the threshold effect hypothesis, in Figure 6 we plot the time series of seller effort for different seller types. More specifically, the two effort measures correspond to the two sets of seller behaviors (i.e., defect and late delivery) under the new regime, as indicated in Table 1, with the horizontal lines indicating the corresponding thresholds. Sellers are grouped by their performance on these measures in the period before the policy announcement. The vertical lines correspond to the policy announcement month and implementation month, respectively. In both graphs, we see a convergence towards the new thresholds for all seller types: sellers who excelled in their effort measures before the policy announcement shirk, sellers who are short of the certification requirements improve their input measures, and both types gravitate towards the threshold after the policy announcement.

Figure 6: Threshold Effect



Notes: The two effort measures correspond to the two sets of seller behaviors under the new regime (shown in Table 1), with the horizontal lines indicating the thresholds. Each curve represent a seller group based on their performance on these measures in the "before" period. The vertical lines correspond to the policy announcement month and implementation month.

We adopt a regression-discontinuity design (RDD) to quantify the threshold effect. Recall that the new certification requirements take into account two metrics: defect rate and late delivery rate. To construct the estimation sample, we find sellers who meet the requirements on defect rate but whose late delivery rate is between 4% and 6% (i.e., within 1% around the 5% threshold) based on the "before" data.[11] We then estimate the threshold effect using the following seller-month level

[11] We focus on sellers who meet the requirements on defect rate but vary in whether they meet the requirements on late delivery rate because the latter is much more binding than the former. We do not have many sellers (an order of magnitude less) to do the analogous RDD where sellers meet the more binding requirement on late delivery rate

regression:

$$Y_{it} = \sum_{q=1}^{q=3} \beta_q * Pre_{q,t} * NotBadgedSim_i + \gamma_1 * Post_{Announce,t} * NotBadgedSim_i + \gamma_2 * Post_{Implement,t} * NotBadgedSim_i + \eta_i + \xi_t + \epsilon_{it}, \quad (1)$$

where Y_{it} is the outcome variable for seller i in month t; $NotBadgedSim_i$ equals 1 if seller i does not meet the new certification requirement on the policy announcement date by simulation; $Post_{Announce,t}$ is the dummy for the months after the policy announcement month; $Post_{Implement,t}$ is the dummy for the months after the policy implementation month; $Pre_{3,t}$ is the dummy for Month -13, -12, and -11; $Pre_{2,t}$ is the dummy for Month -10, -9, and -8; $Pre_{1,t}$ is the dummy for Month -6 and -7; η_i and ξ_t are seller and month fixed effects, respectively.

Our parameters of interest are γ_1 and γ_2, which capture the difference in the temporal changes in effort between (a) sellers that were immediately selected for the eTRS by the new regime upon the announcement of the new policy and (b) those that were not automatically qualified for the eTRS at the policy announcement time. The β's capture any pre-existing differences in the two groups of sellers before the policy announcement. Since the omitted group in the regression is Month -16, -15, and -14, all these estimated coefficients should be interpreted relative to the difference in the baseline outcome in these three months.

Results are reported in Table 3. Column (1) shows that sellers who are simulated not to be badged (i.e., those with a pre-existing average late delivery rate between 5% and 6%) improve on delivery speed relative to sellers who are simulated to be badged (i.e., those between 4% and 5%) after the policy announcement date. This result is consistent with the threshold effect in that, relatively speaking, sellers just below the bar are motivated to exert effort because the marginal benefit of doing so is large, and sellers just above the bar tend to shirk to stay just above the bar. In column (2), we control for $Post_{Implement}$, which measures the additional effect on top of $Post_{Announce,t}$. The positive coefficient estimate suggests that the threshold effect is even stronger after the policy implementation date. In column (3), we add the dummies for the months before the policy announcement and find the insignificant estimates are consistent with the parallel trend assumption between the two seller groups.

In columns (4) to (6), we conduct placebo analyses by estimating equation 1 using the monthly defect rate as the outcome variable. Unlike the results on late delivery rate, we do not see any

but do not meet the less binding requirement on defect rate.

Table 3: Threshold Effect

	(1) late delivery	(2) late delivery	(3) late delivery	(4) defect	(5) defect	(6) defect
NotBadgedSim*$Post_{Announce}$	-0.622***	-0.456***	-0.449***	-0.016	-0.013	-0.005
	(0.042)	(0.052)	(0.070)	(0.015)	(0.017)	(0.018)
NotBadgedSim*$Post_{Implement}$		-0.313***	-0.313***		-0.006	-0.006
		(0.058)	(0.058)		(0.023)	(0.023)
NotBadgedSim*Pre3			0.047			0.002
			(0.067)			(0.007)
NotBadgedSim*Pre2			-0.015			0.016**
			(0.071)			(0.008)
NotBadgedSim*Pre1			-0.009			0.016*
			(0.078)			(0.009)
Observations	1,013,079	1,013,079	1,013,079	1,013,079	1,013,079	1,013,079
R-squared	0.102	0.102	0.102	0.091	0.091	0.091
Seller Fixed Effect	✓	✓	✓	✓	✓	✓
Month Fixed Effect	✓	✓	✓	✓	✓	✓

Notes: One observation is a seller-month pair. Outcome variables are late delivery rate and defect rate in percentage points. The term NotBadgedSim equals 1 if the seller does not meet the new certification requirement on the policy announcement date; $Post_{Announce}$ and $Post_{Implement}$ are the dummies for the months after the policy announcement month and implementation month, respectively; Pre3 is the dummy for Month -13, -12, and -11; Pre2 is the dummy for Month -10, -9, and -8; Pre1 is the dummy for Month -6 and -7. The omitted group in the regression is Month -16, -15, and -14. We also control for the constant term in the regression. Standard errors in parentheses and clustered at the seller level. *** p<0.01, ** p<0.05, * p<0.1.

statistically significant change in sellers' defect rate after the policy announcement. This result provides further support to the threshold effect, which predicts that sellers should have little incentive to further improve on defect rate if they had met the bar on defect rate by the policy announcement date.

Lastly, we study the existence of multitasking, namely whether sellers shirk on consumer-relevant quality dimensions that are not in the new certification requirements. In the spirit of Holmstrom and Milgrom (1991), we test whether focusing on the certification-highlighted quality metrics comes at the cost of deteriorating overall quality using the RDD specification in equation 1. The evidence suggests that sellers do not perform worse in these quality metrics after the policy change (see detailed results in Appendix D).

5 Results on Buyers

5.1 Certification Premium

Our conceptual framework predicts that the change in buyers' valuation towards the eTRS certificate is ambiguous after the regime change. Regardless of the overall change, the regime change should leave consumers to favor the eTRS certificate more in markets where the administrative data is more relevant for consumer experience. To test the hypotheses, we estimate changes in the eTRS premium using the matched listings approach first seen in Einav et al. (2011) and Elfenbein et al. (2012).

Specifically, we match listings by seller ID, listing title and subtitle, Product ID[12], listing price, and listing date. The goal of matching is to control for unobserved product heterogeneity and temporal demand and supply shocks that could be correlated with sales probability and a seller's eTRS status. Within a given matched set, we then compare the sales probability across listings with and without the eTRS badge. Note that there is variation in the eTRS badge across listings even for the same seller on the same day because on eBay, sellers with the *eTRS status* can get the visible *eTRS badge* on a listing only if they offer 1-day handling and 15-day return for that listing. By "eTRS badge" or "badge", we refer to the signal that is observable to consumers. Conceptually, the variation in the eTRS badge conditional on the matching procedure can be thought of as seller experiments, where sellers randomize the appearance of eTRS badge to learn consumer demand (Einav et al. (2011) provides more details for this argument).

The matching procedure yields a sample of 20,341 unique sellers in 313 markets, with more than 3 million listings in our sample period between Nov 20, 2015 and May 19, 2016 (i.e., three months before and after the policy implementation). Our regression equation is specified as follows:

$$Success_{ij(t)} = \beta_1\ badge_{ij(t)} + \beta_2\ badge_{ij(t)} * Post_{Implement,t} + \eta_{j(t)} + \epsilon_{ij(t)}, \tag{2}$$

where $Success_{ij(t)}$ is a dummy for whether listing i (listed on day t) in the matched set j eventually results in a sale; $badge_{ij(t)}$ is the dummy for whether listing i has the eTRS badge; $Post_{Implement,t}$ is the dummy for the months after the policy implementation; $\eta_{j(t)}$ are matched sets fixed effects; $\epsilon_{ij(t)}$ are idiosyncratic errors.

The estimation results are reported in Table 4. The first column shows that consumers are 3.5

[12]Product ID is eBay's finest catalog, which is defined for homogeneous products only. For example, an unlocked 128GB black iPhone 12 has a unique Product ID that is different from that of another version of iPhone.

Table 4: Certification Premium

Outcome variable: Dummy for whether a listing results in sales			
	(1)	(2)	(3)
	all markets	low relevance	high relevance
badge	0.035***	0.053***	0.026***
	(0.001)	(0.001)	(0.001)
badge* $\text{Post}_{Implement}$	0.016***	0.010***	0.020***
	(0.001)	(0.002)	(0.001)
Percentage change	46%	20%	78%
Observations	3,387,161	1,110,474	2,276,687
R-squared	0.087	0.097	0.081
Matched set Fixed Effect	✓	✓	✓

Notes: One observation is a listing. The outcome is whether a listing sells; $Post_{Implement}$ are the dummies for the months after the policy implementation month. Listings are matched based on seller ID, listing title and subtitle, Product ID, listing date, and listing price. Standard errors are clustered at the matched listing level. *** p<0.01, ** p<0.05, * p<0.1.

percentage points more likely to purchase from a listing with the eTRS badge than from an otherwise identical listing without the badge. Also, the eTRS premium is higher after the policy change. We should be careful in concluding that the eTRS premium increased because the higher premium may reflect temporal variation in badge premium moving from holiday to non-holiday seasons. However, assuming that the time trend is the same for markets with high vs. low relevance, we can compare consumers' valuation of the quality certificate across markets. For the clarity of presentation, in the main text we focus on market comparison by relevance (rather than by relevance and controllability), because relevance is more closely related to what consumers care about in quality certification. The same analyses of certification premium by the four market types defined by relevance-by-controllability can be found in Appendix E.

Comparing columns 2 and 3 leads to two findings. First, before the policy change, the badge premium is about 2.7 percentage points (pp) higher in markets with low relevance than in markets with high relevance (5.3 pp - 2.6 pp). As shown in Table 2, low-relevance markets tend to be those with less standardized products, and consumers should value ratings (and thus quality certificates) more in these markets because quality cannot be easily reviewed through pictures and item descriptions. After the policy change, however, the increase in certification premium is higher in high-relevance markets, both in absolute terms (2.0 vs. 1.0 pp) and in relative terms with respect to the before period (78% vs. 20%). The p-value of the Wald test on the difference in the percentage increase (i.e., 78% and 20%) is less than 0.001. This result suggests that the change

in certification premium is relatively more positive in markets where administrative data on seller behavior is more aligned with consumer reports.

As a robustness check, we re-run Equation 2 to identify an alternative definition of the eTRS premium — the difference in equilibrium sales price for otherwise-similar transactions that differ only in the eTRS badge. This eTRS premium in terms of price difference reflects the consumer's willingness to pay for the eTRS badge in the market equilibrium. Similar to the matching procedure detailed previously, we match *transactions* by seller ID, listing title and subtitle, Product ID, and transaction date. That is, the transactions within a given matched set can differ only in the eTRS badge and possibly the outcome (i.e., the sales price). The matching procedure yields a sample of 26,079 transactions belonging to 4,304 unique matched sets, covering 1,334 unique sellers and 224 unique product markets between Nov 20, 2015 and May 19, 2016 (i.e., three months before and after the policy implementation). Similar to the results above, we find that the eTRS badge warrants an overall positive price premium. In addition, the price premium for the badge becomes larger (in relative terms) in high-relevance markets than in low-relevance markets after the policy change, suggesting a higher consumer willingness to pay in markets where the new certification reflects more of seller qualities valued by consumers. The regression table and detailed discussions are provided in Appendix E.

5.2 Consumer Retention

Having studied the change in consumers' valuation of the badge, we now study the downstream consequence of using administrative-data-based quality certification in terms of consumer retention. Our sample is as follows: starting from all transactions in our raw data set (described in Section 2), we construct a buyer-market-month level data set. For each observation, we then define the outcome variable as a dummy for whether the buyer purchases another item on eBay (in any market) within the six months after the transaction month of the focal purchase. Lastly, we aggregate the sample to the market-month level by averaging this dummy variable across buyers in each market-month. Our regression is specified as follows:

$$Y_{mt} = \beta_1 * Relevance_m * Post_{Implement,t} + \beta_2 * Controllability_m * Post_{Implement,t} + \eta_{m,\tilde{t}} + \xi_t + \epsilon_{mt}, \quad (3)$$

where Y_{mt} is the average retention rate in market m in month t; $Relevance_m$ and $Controllability_m$ are dummies for whether market m has above-median relevance or controllability; ξ_t is market and

month fixed effects; $\eta_{m,\tilde{t}}$, $\tilde{t} \in \{1, 2, 3, 4\}$ is market-specific quarter fixed effects, which control for different seasonality in markets; lastly, ϵ_{mt} are idiosyncratic errors.

The results are reported in Table 5. Column 1 shows that in markets where input measures are more relevant, consumers are 0.2 percentage points (or about 0.24% over the baseline) more likely to make another purchase on eBay within six months after the focal purchase. As reported in column 2, we find essentially the same result after incorporating the controllability dummy in the regression. While the magnitude of the estimates is small, the sign of the coefficient is consistent with our conceptual framework: in markets where administrative data on seller behavior is more aligned with consumer reports, consumers are more likely to come back to the platform and purchase again in the future. This result confirms that consumers value the eTRS badge more in the markets where the new, administrative data-based eTRS criteria are better aligned with consumer experience.

Table 5: Consumer Retention

Outcome Variable: retention rate in 6 months		
	(1)	(2)
Relevance*Post$_{Implement}$	0.002*	0.002*
	(0.001)	(0.001)
Control*Post$_{Implement}$		0.000
		(0.001)
Constant	0.842***	0.842***
	(0.000)	(0.000)
Observations	7,169	7,169
R-squared	0.970	0.970
Market-Seasonality Fixed Effect	✓	✓
Month Fixed Effect	✓	✓

Notes: One observation is a buyer-market-month pair. Outcome variable is whether the consumer buys again in the following six months in any market; $Post_{Implement}$ are the dummies for the months after the policy implementation month. Standard errors in parentheses and clustered at the market level. *** p<0.01, ** p<0.05, * p<0.1.

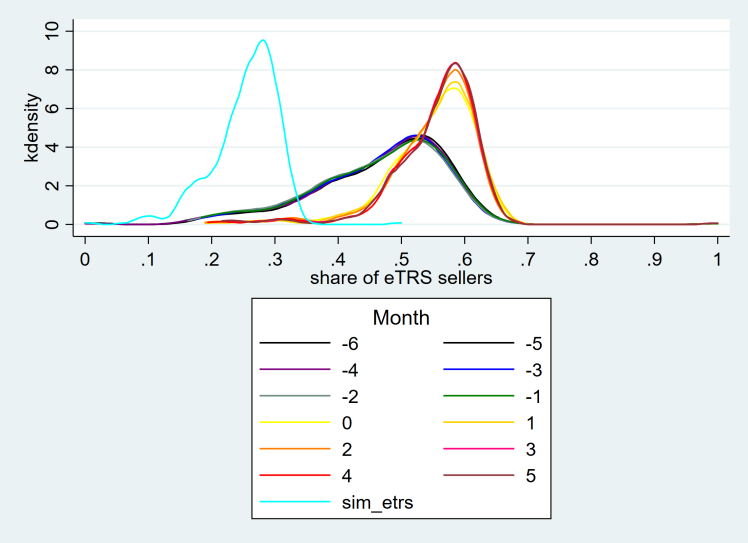
6 Results on Market Outcomes

6.1 Homogenization of Share of Certified Sellers across Markets

Because the policy change increases seller controllability of certification management, there should be a homogenization in the share of certified sellers across markets after the policy change. In

fact, Figure 5 before shows consistent evidence that this is the case: fewer sellers are certified in critical markets before the regime change, but the simulated shares are much more similar across the two types of markets, and the "after" shares of certification are essentially identical except for the largest sellers. This observation suggests that the homogenization result is driven by both the selection effect and seller behavioral change.

Figure 7: Share of Certified Sellers across Markets



Notes: Curves with cool-toned (resp., warm-toned) colors are for months before (resp., after) the policy change. The light blue curve corresponds to simulated eTRS status, normalized by a constant.

To provide more direct evidence on this, in Figure 7 we plot the distribution of the share of certified sellers across all 336 markets. The six curves with cool-toned colors (the bunch of curves in the middle of the graph) represent the share of sellers with actual eTRS status in the months before the regime change (Month -6 to -1), the curves with warm-toned colors (the bunch of curves on the right of the graph) represent the share of sellers with actual eTRS status in the months after the regime change (Month 0 to Month 5), and the light blue curve (the curve on the left of the graph) corresponds to the density of the simulated eTRS status upon policy announcement (Month -5) normalized by a constant. A key takeaway from the figure is that both the cross-market distributions of the actual eTRS share and the simulated eTRS share become more concentrated when the new certification is based mainly on administrative data. Depending on the context and the goal of the market regulator, the homogenization of certified sellers across markets may be a desirable result because it could help consumers, especially less experienced ones, to better interpret

the eTRS signal.[13]

6.2 Sales and Market Concentration

So far, we have established that changes in certification requirements can affect the selection and behavior of sellers, in turn affecting buyer expectation about seller quality. A different expectation on seller quality could affect different seller types differently and, therefore, change sales concentration within each market. As we have seen before, sellers in markets with different controllability change their behavior differently after the regime change, and buyers' perception of the quality certificate depends on how relevant the administrative data is in a market. What are the market-level effects of switching from using mainly consumer reports to administrative data in quality certification?

To answer this question, we use the following DiD specification at the market-month level:

$$Y_{mt} = \beta_1 * lrlc_m * Post_t + \beta_2 * hrhc_m * Post_t + \beta_3 * hrlc_m * Post_t + \eta_{m,\tilde{t}} + \xi_t + \epsilon_{mt}, \qquad (4)$$

where Y_{mt} are outcomes in market m in month t; $Post_t$ is the dummy for whether month t is after the policy implementation date; $lrlc_m$ indicates whether market m is "low-relevance-low-controllability"; Similarly, $hrhc_m$ and $hrlc_m$ indicate whether a market is "high-relevance-high-controllability" and "high-relevance-low-controllability", respectively; ξ_t is month fixed effects. We also use $\eta_{m,\tilde{t}}$, $\tilde{t} \in \{1, 2, 3, 4\}$, which is market-specific quarter fixed effects, to control for different seasonality in markets.[14] All regressions cluster standard errors by market. The times series of the market outcomes are reported in Figure 11 in the Appendix F.

Results are reported in Table 6. In columns (1) and (2), the outcome variables are the natural log of sales in USD and in quantity. There are no statistically significant differences in these variables across different types of markets after the policy implementation.

Next, we study how market concentration changes after the policy implementation. The outcome variables in columns (3) and (4) are the natural log of the number of sellers who have any sales and the natural log of HHI (potentially ranges from 0 to 10,000) based on dollar sales per seller in our sample. We see that after the policy implementation, in markets with high relevance and low

[13] The homogenization result may not always be desirable: to the extent that the different distributions of certified sellers reflect the difference in the underlying distribution of sellers (e.g., it is easier to sell electronics than art), then market designers may want to see different shares of badged sellers across markets.

[14] As a side note, in the previous analyses on sellers in Section 4 and on buyers in Section 5, we focus on studying heterogeneity by either controllability or relevance, respectively. In the current market-level analyses, we focus on the four types of markets by controllability-by-relevance, because both supply-side and demand-side response can affect market outcomes.

Table 6: Sales and Market Concentration

	(1) log(sales volume USD)	(2) log(sales quantity)	(3) log(number sellers with any sales)	(4) log(HHI)	(5) share sellers with sales who are small	(6) share sales (USD) from small sellers
$\text{lrlc\_Post}_{Implementation}$	-0.148	-0.110	-0.126	0.042	-0.008***	-0.006
	(0.206)	(0.146)	(0.116)	(0.044)	(0.002)	(0.008)
$\text{hrhc\_Post}_{Implementation}$	-0.015	-0.022	-0.014	0.006	-0.005**	0.002
	(0.038)	(0.025)	(0.014)	(0.048)	(0.002)	(0.009)
$\text{hrlc\_Post}_{Implementation}$	-0.065	-0.063	-0.100*	0.134***	-0.010***	-0.008
	(0.110)	(0.076)	(0.057)	(0.042)	(0.002)	(0.008)
Constant	14.048***	10.398***	7.995***	4.548***	0.768***	0.511***
	(0.015)	(0.011)	(0.008)	(0.007)	(0.000)	(0.001)
Observations	7,435	7,435	7,435	7,406	7,406	7,406
R-squared	0.868	0.929	0.936	0.949	0.984	0.969
Market-Seasonality FE	✓	✓	✓	✓	✓	✓
Month FE	✓	✓	✓	✓	✓	✓

Notes: One observation is a market-month pair; Post is the dummy for transactions after the policy implementation date; $lrlc_m$ indicates whether market m is "low-relevance-low-controllability"; Similarly, $hrhc_m$ and $hrlc_m$ indicate whether a market is "high-relevance-high-controllability" and "high-relevance-low-controllability", respectively. We use market-quarter fixed effects (FE) to control for market seasonality. HHI ranges from 0 to 10,000. Observations with top and bottom 1 percentiles of outcome variables removed. In parentheses are standard errors clustered at the market level. *** p<0.01, ** p<0.05, * p<0.1.

controllability (hrlc), there is a decrease in the share of sellers who successfully made a sale and an increase in HHI, relative to the benchmark (markets of low relevance and high controllability). Lastly, we study changes in the market share of small sellers. Column (5) shows that out of all sellers with positive sales, the proportion of small sellers declines in markets with less controllability. In column (6), the outcome variable is the share of sales (in dollars) that comes from small sellers. These two estimates suggest that small sellers are less likely to sell and have smaller market shares in markets with high relevance and low controllability after the regime change.

Taken together, these results suggest that markets with high relevance and low controllability become more concentrated towards big sellers after the policy implementation. These are markets in which both sellers and buyers are relatively better off after the policy change: sellers benefit because the low controllability in consumer reports (and therefore the eTRS certificate) is greatly reduced by the regime change; buyers benefit because they appreciate the high relevance of the new eTRS certificate. In these markets, one would expect such a regime change to benefit small sellers more, because in the old regime randomness in consumer reports is less likely to cancel out in the average metrics for small sellers than for big sellers, i.e., the law of large numbers hasn't kicked

in for smaller sellers. However, the empirical evidence suggests that the exact effect on market concentration may critically depend on the specific market institutions. In our setting, sellers gain the eTRS status at the seller level and a seller needs to have a minimum number of past sales to be eligible for the eTRS status. Both rules make the incentive to improve effort to meet the certification requirements stronger for large sellers. The results on market concentration may differ in other marketplaces where these two rules are different.

7 Conclusion and Discussion

In this paper, we study the effect of changing from using mainly consumer reports to administrative data in quality certification on a leading e-commerce platform. We find that the policy change motivates sellers to exert effort in dimensions highlighted in the new certificate, but it also induces sellers to more precisely target their effort at the minimum quality threshold. Additionally, in markets where the new quality certificate is more aligned with consumer experience, buyers place a higher premium on the quality certificate and are more likely to purchase from the platform again in the following six months. Lastly, the policy change homogenizes the share of certified sellers across markets, and sales are more concentrated towards large sellers in markets with lower controllability.

These results yield a few lessons for market designers who consider the use of consumer reports and administrative data in constructing quality certification. First, the key trade-off is that incorporating consumer reports into certification makes it more relevant for the ultimate consumer experience, but doing so may discourage seller effort because consumer-reported information can be driven by random factors not entirely within sellers' control. Therefore, choosing the information types in quality certification depends on market characteristics and welfare weights placed on consumers versus sellers. Specifically, if product quality can be measured accurately with administrative data on seller behavior, market designers should avoid using consumer reports. One example is the application programming interfaces (APIs) for data feed, such as an API that gives real-time exchange rate between U.S. dollars and euros. The quality of these APIs can be captured in terms of the accuracy of data feed (e.g., percentage deviation from the ground truth) and the latency of the feed. In this example, adding consumer reports would introduce randomness to service providers without a significant benefit. However, in markets where administrative data on seller behavior is a poor predictor of seller quality, such as the market for nannies, market designers should put less weight on this data (such as whether the nanny arrives at work on time), but more

weight on consumer reports. Lastly, if market designers place more welfare weight on consumers (resp., sellers), then they should increase the weight of consumer reports (resp., administrative data on seller behavior) in the construction of quality certificates.

Second, given the discrete nature of certification thresholds, reducing the weight on consumer reports may affect seller effort in ambiguous directions – on the one hand, a bigger emphasis on administrative data on seller behavior enhances seller controllability and motivates sellers to improve the focal metrics; on the other hand, better controllability motivates sellers to target the quality threshold and stop improving beyond the threshold. Therefore, market designers that aim to encourage seller effort overall may want to choose an optimal level of randomness in the certification threshold. While there has been a growing theoretical literature on optimal information control (e.g., Vellodi (2018), Saeedi and Shourideh (2019), Shi et al. (2020)), there has not been a study on the optimal design of noise in certification thresholds. Our empirical results suggest that this design parameter may deserve more explicit attention in future research.

Third, our results show that using administrative data in quality certification results in a homogenization of the share of certified sellers across markets. Depending on empirical settings, market designers may find this feature desirable, because a more even distribution across different product markets could help consumers understand the certification signal. However, if different distributions of certified sellers reflect the underlying distributions of seller quality in different markets, then this homogenization may not be desirable.

Lastly, our results suggest that the design of quality certificates can have long-run implications on the marketplace because it can potentially change market concentration. The direction of the change depends on market institutions, and in particular, whether the certification is at the seller level or at the listing level, the certification's requirement on minimum number of past sales, and any other institutions that could disproportionately affect the benefit and cost of certification for certain groups of sellers.

Our study is subject to a few limitations. First, the analysis is based on a single marketplace that embodies simultaneously a seller reputation system (seller ratings based on consumer reports), a seller quality certification program (eTRS), and buyer warranty. This feature implies that (1) some experienced consumers may always observe certain consumer reports even after these reports were removed from the criteria of quality certification[15] and (2) eBay's generous buyer warranty program reduces the impact of quality certificates (Hui et al. (2016)). With these caveats in mind,

[15]See Section 2 for our argument on why this feature is unlikely a big problem empirically in the seting of eBay.

however, we note that multifaceted trust systems are common in e-commerce. For example, not only does Amazon report product-level and seller-level ratings based on consumer reports, but it also highlights the Amazon's Choice badge and offers strong buyer warranty. The Superhost badge on Airbnb requires a consumer rating no lower than 4.8, and the Top Rated badge on Upwork requires a 90% or higher success rate; in addition, both Airbnb and Upwork publicize consumer reports separately from these badges. These examples suggest that our results are likely relevant for other digital platforms.

The second limitation of our study is that, because all sellers in our sample are subject to the regime change, we do not have a clean control group. We can mainly conduct event-study-style analyses and comparisons across different types of markets. Nevertheless, we provide one of the first evidence on the trade-offs of different ways of measuring quality in quality certification, which is ubiquitously used in everyday life (Dranove and Jin, 2010) but not much studied in the literature.

Lastly, our paper cannot quantify welfare changes due to the policy change, which would require a structural model. That said, in terms of consumer welfare, the result on increased badge premium suggests that consumers continue to value the information in the eTRS badge. We also cannot speak to changes in seller surplus or total welfare, as we do not observe the costs of seller effort. Although we observe a relative increase in sales concentration in markets with low controllability after the policy change, this increase may be welfare enhancing because large sellers may be more responsive to the eTRS-highlighted incentives and more cost-efficient in effort improvement. Finally, we do not have any information on consumer search, which could change in light of the new certification system and how the certificate is incorporated in the platform's search algorithm. The role of certification in consumer search is a topic that warrants future research.

References

Alé-Chilet, Jorge and Sarah Moshary, "Beyond consumer switching: Supply responses to food packaging and advertising regulations," *Marketing Science*, 2022, *41* (2), 243–270.

Anderson, Eric T and Duncan I Simester, "Reviews without a purchase: Low ratings, loyal customers, and deception," *Journal of Marketing Research*, 2014, *51* (3), 249–269.

Anderson, Eugene W and Mary W Sullivan, "The antecedents and consequences of customer satisfaction for firms," *Marketing science*, 1993, *12* (2), 125–143.

Bai, Jie, "Melons as lemons: Asymmetric information, consumer learning and quality provision," Technical Report, Working paper 2018.

Baker, George P, "Incentive contracts and performance measurement," *Journal of political Economy*, 1992, *100* (3), 598–614.

Barach, Moshe A, Joseph M Golden, and John J Horton, "Steering in online markets: the role of platform incentives and credibility," *Management Science*, 2020.

Benson, Alan, Aaron Sojourner, and Akhmed Umyarov, "Can reputation discipline the gig economy? Experimental evidence from an online labor market," *Management Science*, 2020, *66* (5), 1802–1825.

Bollinger, Bryan, Phillip Leslie, and Alan Sorensen, "Calorie posting in chain restaurants," *American Economic Journal: Economic Policy*, 2011, *3* (1), 91–128.

Bolton, Gary, Ben Greiner, and Axel Ockenfels, "Engineering trust: reciprocity in the production of reputation information," *Management science*, 2013, *59* (2), 265–285.

Brandes, Leif, David Godes, and Dina Mayzlin, "What drives extremity bias in online reviews? Theory and experimental evidence," *Theory and Experimental Evidence (September 6, 2019)*, 2019.

Burtch, Gordon, Yili Hong, Ravi Bapna, and Vladas Griskevicius, "Stimulating online reviews by combining financial incentives and social norms," *Management Science*, 2018, *64* (5), 2065–2082.

Cabral, Luis and Ali Hortacsu, "The dynamics of seller reputation: Evidence from eBay," *The Journal of Industrial Economics*, 2010, *58* (1), 54–78.

_ and Lingfang Li, "A dollar for your thoughts: Feedback-conditional rebates on eBay," *Management Science*, 2015, *61* (9), 2052–2063.

Chevalier, Judith A and Dina Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research*, 2006, *43* (3), 345–354.

Dai, Weijia, Ginger Z Jin, Jungmin Lee, and Michael Luca, "Aggregation of consumer ratings: an application to yelp.com," *Quantitative Marketing and Economics*, 2018, *16* (3), 289–339.

Dellarocas, Chrysanthos, "The digitization of word of mouth: Promise and challenges of online feedback mechanisms," *Management science*, 2003, *49* (10), 1407–1424.

_ and Charles A Wood, "The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias," *Management science*, 2008, *54* (3), 460–476.

_ , Ritu Narayan et al., "A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth," *Statistical science*, 2006, *21* (2), 277–285.

Dewan, Sanjeev and Vernon Hsu, "Adverse selection in electronic markets: Evidence from online stamp auctions," *The Journal of Industrial Economics*, 2004, *52* (4), 497–516.

Dranove, David and Ginger Zhe Jin, "Quality disclosure and certification: Theory and practice," *Journal of Economic Literature*, 2010, *48* (4), 935–63.

Einav, Liran, Chiara Farronato, and Jonathan Levin, "Peer-to-peer markets," *Annual Review of Economics*, 2016, *8*, 615–635.

_ , Theresa Kuchler, Jonathan D Levin, and Neel Sundaresan, "Learning from seller experiments in online markets," Technical Report, National Bureau of Economic Research 2011.

Elfenbein, Daniel W, Ray Fisman, and Brian McManus, "Charity as a substitute for reputation: Evidence from an online marketplace," *Review of Economic Studies*, 2012, *79* (4), 1441–1468.

_ , Raymond Fisman, and Brian McManus, "Market structure, reputation, and the value of quality certification," *American Economic Journal: Microeconomics*, 2015, *7* (4), 83–108.

Farronato, Chiara, Andrey Fradkin, Bradley Larsen, and Erik Brynjolfsson, "Consumer Protection in an Online World: An Analysis of Occupational Licensing," Technical Report, National Bureau of Economic Research 2020.

Filippas, Apostolos, John J Horton, and Joseph M Golden, "Reputation inflation," *Marketing Science*, 2022.

Fradkin, Andrey and David Holtz, "More Reviews May Not Help: Evidence from Incentivized First Reviews on Airbnb," *arXiv preprint arXiv:2112.09783*, 2021.

_ , Elena Grewal, and David Holtz, "Reciprocity in Two-sided Reputation Systems: Evidence from an Experiment on Airbnb," 2019.

_ , _ , and _ , "Reciprocity and unveiling in two-sided reputation systems: Evidence from an experiment on airbnb," *Marketing Science*, 2021, *40* (6), 1013–1029.

He, Sherry, Brett Hollenbeck, and Davide Proserpio, "The market for fake reviews," *Available at SSRN*, 2020.

Ho, Yi-Chun, Junjie Wu, and Yong Tan, "Disconfirmation Effect on Online Rating Behavior: A Structural Model," *Information Systems Research*, 2017, *28* (3), 626–642.

Hollenbeck, Brett, Sridhar Moorthy, and Davide Proserpio, "Advertising strategy in the presence of reviews: An empirical analysis," *Marketing Science*, 2019, *38* (5), 793–811.

Holmstrom, Bengt and Paul Milgrom, "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design," *JL Econ. & Org.*, 1991, *7*, 24.

Hu, Nan, Paul A Pavlou, and Jie Jennifer Zhang, "Why do online product reviews have a J-shaped distribution? Overcoming biases in online word-of-mouth communication," *Communications of the ACM*, 2009, *52* (10), 144–147.

Hui, Xiang, Maryam Saeedi, and Neel Sundaresan, "Adverse Selection or Moral Hazard, An Empirical Study," *The Journal of Industrial Economics*, 2018, *66* (3), 610–649.

_ , _ , Giancarlo Spagnolo, and Steve Tadelis, "Certification, Reputation and Entry: An Empirical Analysis," *Unpublished Manuscript*, 2017.

_ , _ , Zeqian Shen, and Neel Sundaresan, "Reputation and regulations: evidence from eBay," *Management Science*, 2016, *62* (12), 3604–3616.

_ , Tobias J Klein, and Konrad O Stahl, "When and Why Do Buyers Rate in Online Markets?," 2022.

_ , Zekun Liu, and Weiqing Zhang, "Mitigating the Cold-start Problem in Reputation Systems: Evidence from a Field Experiment," *Available at SSRN*, 2020.

Hunter, Megan, "Chasing Stars: Firms' Strategic Responses to Online Consumer Ratings," *Available at SSRN 3554390*, 2020.

Ishihara, Masakazu and Yuzhou Liu, "A Dynamic Structural Model of Endogenous Consumer Reviews in Durable Goods Markets," *Available at SSRN 2728524*, 2017.

Jin, Ginger Z., Zhentong Lu, Xiaolu Zhou, and Chunxiao Li, "The Effects of Government Licensing on E-commerce: Evidence from Alibaba," *Journal of Law & Economics*, 2022.

Jin, Ginger Zhe and Phillip Leslie, "The effect of information on product quality: Evidence from restaurant hygiene grade cards," *The Quarterly Journal of Economics*, 2003, *118* (2), 409–451.

Karaman, Hülya, "Online review solicitations reduce extremity bias in online review distributions and increase their representativeness," *Management Science*, 2021, *67* (7), 4420–4445.

Klein, Tobias J, Christian Lambertz, and Konrad O Stahl, "Market transparency, adverse selection, and moral hazard," *Journal of Political Economy*, 2016, *124* (6), 1677–1713.

Leland, Hayne E., "Quacks, Lemons and Licensing: a Theory of Minimum Quality Standards," *Journal of Political Economy*, 1979, *87* (6), 1328–1346.

Lewis, Gregory and Georgios Zervas, "The welfare impact of consumer reviews: A case study of the hotel industry," *Unpublished manuscript*, 2016.

Li, Lingfang Ivy, Steven Tadelis, and Xiaolan Zhou, "Buying reputation as a signal of quality: Evidence from an online marketplace," *RAND Journal of Economics*, 2022.

Luca, Michael, "Reviews, reputation, and revenue: The case of Yelp. com," *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, 2016, (12-016).

_ **and Georgios Zervas**, "Fake it till you make it: Reputation, competition, and Yelp review fraud," *Management Science*, 2016, *62* (12), 3412–3427.

_ **and Oren Reshef**, "The effect of price on firm reputation," *Management Science*, 2021, *67* (7), 4408–4419.

Marinescu, Ioana, Andrew Chamberlain, Morgan Smart, and Nadav Klein, "Incentives can reduce bias in online employer reviews.," *Journal of Experimental Psychology: Applied*, 2021.

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier, "Promotional reviews: An empirical investigation of online review manipulation," *American Economic Review*, 2014, *104* (8), 2421–55.

Moe, Wendy W and Michael Trusov, "The value of social dynamics in online product ratings forums," *Journal of Marketing Research*, 2011, *48* (3), 444–456.

Nosko, Chris and Steven Tadelis, "The limits of reputation in platform markets: An empirical analysis and field experiment," Technical Report, National Bureau of Economic Research 2015.

Pallais, Amanda, "Inefficient hiring in entry-level labor markets," *American Economic Review*, 2014, *104* (11), 3565–99.

Park, Sungsik, Woochoel Shin, and Jinhong Xie, "The fateful first consumer review," *Marketing Science*, 2021.

Reimers, Imke and Joel Waldfogel, "Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings," *American Economic Review*, 2021, *111* (6), 1944–71.

Resnick, Paul, Richard Zeckhauser, John Swanson, and Kate Lockwood, "The value of reputation on eBay: A controlled experiment," *Experimental economics*, 2006, *9* (2), 79–101.

Saeedi, Maryam, "Reputation and adverse selection, theory and evidence from eBay," *RAND Journal of Economics*, 2019.

_ **and Ali Shourideh**, "Optimal Rating Design," 2019.

Shapiro, Carl, "Premiums for high quality products as returns to reputations," *Quarterly journal of economics*, 1983, *98* (4), 659–679.

Shi, Zijun June, Kannan Srinivasan, and Kaifu Zhang, "Design of Platform Reputation Systems: Optimal Information Disclosure," 2020.

Tadelis, Steven, "Reputation and feedback systems in online platform markets," *Annual Review of Economics*, 2016, *8*, 321–340.

Vana, Prasad and Anja Lambrecht, "The effect of individual online reviews on purchase likelihood," *Marketing Science*, 2021.

Vatter, Benjamin, "Quality disclosure and regulation: Scoring design in medicare advantage," Technical Report, Working Paper 2021.

Vellodi, Nikhil, "Ratings design and barriers to entry," *Available at SSRN 3267061*, 2018.

Wu, Chunhua, Hai Che, Tat Y Chan, and Xianghua Lu, "The economic value of online reviews," *Marketing Science*, 2015, *34* (5), 739–754.

Zhu, Feng and Xiaoquan Zhang, "Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics," *Journal of marketing*, 2010, *74* (2), 133–148.

Appendix A Seller Size

In this section, we elaborate why seller size matters for the effect of the regime change from using consumer reports to administrative data on seller behavior in quality. Specifically, we provide a numerical simulation to illustrate this point.

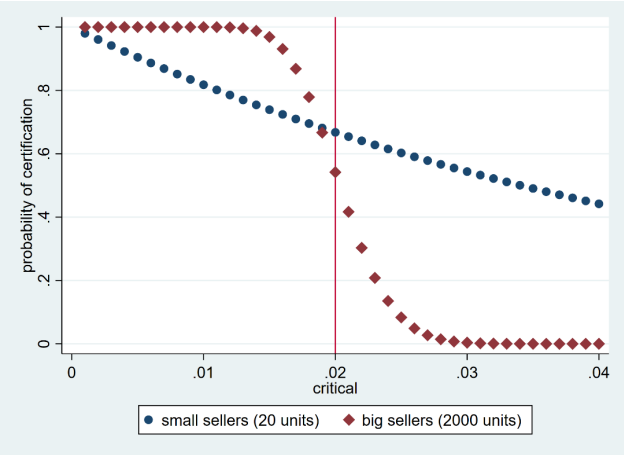
In the old regime, using consumer reports for quality certification present randomness to sellers in that consumer reports are only partially determined by seller effort. A major source of the variance of the randomness is seller size, essentially because all eTRS-relevant metrics are averaged across a seller's qualifying orders, and the law of large numbers applies only for large sellers. More specifically, under the assumption that the orders of a given seller are independent draws, the mean of this randomness may be the same for different sellers in the same category, but a seller with more sales will have a more *precise* average metric relative to her true quality.

To illustrate how the mean and variance of the randomness affect the certificate signal in the old eTRS system, we conduct a numerical simulation where every seller makes maximum effort in the actual performance (say, sellers always describe the item perfectly, always ship immediately, and never cancel), but a seller earns an eTRS certificate if and only if the defect rate observed by eBay is no higher than 2%. The signal reflects an average of previous consumer reports on the seller, and each consumer report conforms to a Bernoulli distribution with probability p for value 1 (i.e., a bad consumer report) and $1-p$ for value 0 (i.e., a good consumer report). One can think of these binary signals as low DSR ratings on the seller-profile page, or claims that consumers file to eBay. Given this, the seller's probability of getting the eTRS badge follows a binomial distribution.

To reflect different controllability in different product markets, we assume p is fixed in each category but ranges from 0 to 1 across categories. Regarding seller size, we consider two types of sellers in each product category: a small seller has only 20 orders qualified for eTRS calculation, while a big seller has 2000 qualified orders. For each type of sellers in a category with p, we simulate the probability of being certified for a typical small seller and a typical big seller in that category. Figure 8 presents the simulation results. The horizontal axis is p; a larger p indicates lower controllability from the seller's perspective. The vertical axis is the probability that a seller satisfies the old eTRS requirement. The vertical line indicates the 2% cutoff. We present two curves, for small and large sellers, respectively.

Figure 8 demonstrates two patterns. First, the probability of getting certified decreases by p (or equivalently, increases by seller controllability) for all sellers. That is, when the draw of the

Figure 8: Seller Size and Controllability



Notes: This figure plots the probability of being certified against p (i.e., lack of seller controllability) for a typical small seller and a typical big seller. Sellers always deliver high quality, and p is the probability of getting negative feedback. The certification bar is at 98% positive feedback, as represented by the vertical line. The probabilities are calculated using the binomial probability mass function.

randomness is more negative on average, the certificate signal is more negative. Second, when consumer reports are more accurate than the certificate cutoff ($p < 2\%$, the left side of the vertical line), small sellers are more vulnerable to large p (i.e., lack of controllability) than large sellers. Essentially, the law of large numbers guarantees that a perfect-performing large seller will almost always pass the minimum threshold, but chances play a bigger adverse role for sellers with only a few transactions. This pattern is reversed when p is larger than the certificate cutoff ($p > 2\%$, the right side of the vertical line), because it is easier for a small seller to have enough lucky draws to get above the certificate cutoff.

What does the regime change mean in our simulation? Because the new eTRS system excludes consumer feedback, low DSR, and resolved buyer claims, it essentially lowers p to p_0 for all markets, where p_0 reflects other randomness that remains in the new system (e.g., cancellation due to factors out of seller control or traffic jam from the seller's warehouse to the postal office). This implies that we should observe less heterogeneity across markets with high and low controllability, conditional on the same seller performance. However, as long as p_0 is positive, heterogeneity across small and big sellers still exists, and how that heterogeneity changes depends on where p and p_0 are in reality. If p is to the left of the certification threshold, say around 1%, then a reduction to p_0 would disproportionately benefit small sellers, as they were more adversely affected by the old eTRS rules than were big sellers. If p is to the right of the certification threshold, say around 3%, then a

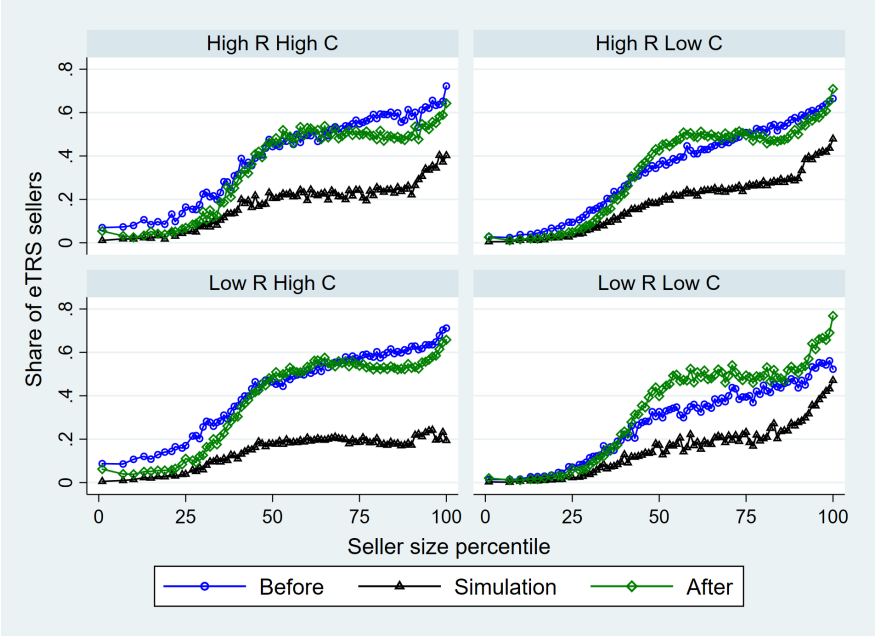
reduction to p_0 would benefit big sellers, because they are almost never certified under the old eTRS rules.

In short, the simulation emphasizes the importance of the interaction between seller controllability in a market and seller size on the selection effect: the regime change should immediately benefit all sellers in markets with low controllability; additionally, it should disproportionately benefit small sellers in markets with low controllability assuming that p is to the left of the certification threshold, i.e., the randomness in consumer reports is not too big. Lastly, small and large sellers may become more homogenized in the share of certified sellers after the regime change.

Appendix B Results on Sellers - Four Types of Markets

In this section, we replicate key results on the selection effect and change in seller behavior by the four types of markets, which differ in both controllability and relevance. We find that controllability is indeed the most important margin for sellers.

Figure 9: Selection Effect and Seller Effort Change, by Market Types, by Seller Size



Notes: Market types are defined based on the median split on relevance and controllability. We shift all curves by a constant (i.e., the same constant for all sub-figures). "Before" refers to Month -5; simulation is done at Month -5. "After" refers to Month 5.

Specifically, similar to Figure 5, we plot in Figure 9 the share of eTRS sellers under the old regime (Month -5), under simulation (using performance metrics up till Month -5), and under

the new regime (Month 5), separately for the four market types defined by the median split of controllability and relevance, and by seller size. The selection effect, defined as the drop from the "before" curve to the simulation curve, is on average smaller in "High R Low C" (high-relevance-low-controllability) markets and "Low R Low C" (low-relevance-low-controllability) markets compared to the other markets. Similarly, the seller effort change, as defined by the increase from the simulation curve to the "after" curve, is on average smaller in "High R Low C" markets and "Low R Low C" markets compared to the other markets. Therefore, the patterns here suggest that compared to relevance, controllability is the seller-facing metric that largely dictates the selection effect and effort change on the seller side.

Now, we quantify the selection effect via the following seller-level regression:

$$Y_i = \beta_0 + \beta_1 * Large_i + \beta_2 * AvgControllability_i + \beta_3 * Large_i * AvgControllability_i + \epsilon_i, \quad (5)$$

where i denotes a seller, Y_i is a categorical variable describing the difference between the seller's simulated and actual certification status in Month -5: it is equal to 1 if a non-certified seller gains a simulated eTRS because of the new eTRS algorithm, -1 if a certified seller loses the eTRS in simulation, and 0 if the simulation does not generate any status change. In the table, $Large_i$ is a dummy indicating whether seller i had at least 400 transactions in Months -8 to -6; $AvgControllability_i$ indicates the degree to which seller i operates in markets with high controllability: it is constructed based on the controllability in each market that seller i operated in during Months -8 to -6, weighted by the seller's sales share in that market in Months -8 to -6. Note that equation 5 is at the seller level because a seller's eTRS status is evaluated at the seller level (not product or listing level). Since we focus on the selection effect at Month -5, the estimation is based on a cross-section of 380,978 sellers, following the sample construction procedure described in Section 3.

Results are reported in Table 7, and we do not report the constant term, to comply with eBay's data policy. From column (1), we see that on average, the net selection effect is 0.078 more negative for large sellers, and 0.152 more negative for sellers in markets with higher controllability (relative to the scale of the dependent variable from -1 to 1). In column (2), we further include an interaction of large seller and markets with average controllability. The negative coefficient on this interaction suggests an extra negative selection for large sellers, relative to small sellers, in markets with higher controllability. Overall, these findings are consistent with the patterns in Figure 5.

Table 7: Seller Selection

	(1)	(2)
	net selection	net selection
Large	-0.077***	-0.061***
	(0.002)	(0.004)
Avg. controllability in operation markets	-0.152***	-0.149***
	(0.001)	(0.001)
Large * Avg. controllability in operation markets		-0.030***
		(0.005)
Observations	424,607	424,607
R-squared	0.030	0.030

Notes: Seller-level cross-sectional regressions. Outcome is the difference between a seller's simulated and actual certification status on the policy announcement date. Large is a dummy for having sold at least 400 items in the three months before the policy announcement. Average controllability is a sales-weighted measure of seller controllability in the markets that a seller operates in. We also control for the constant term in the regression. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

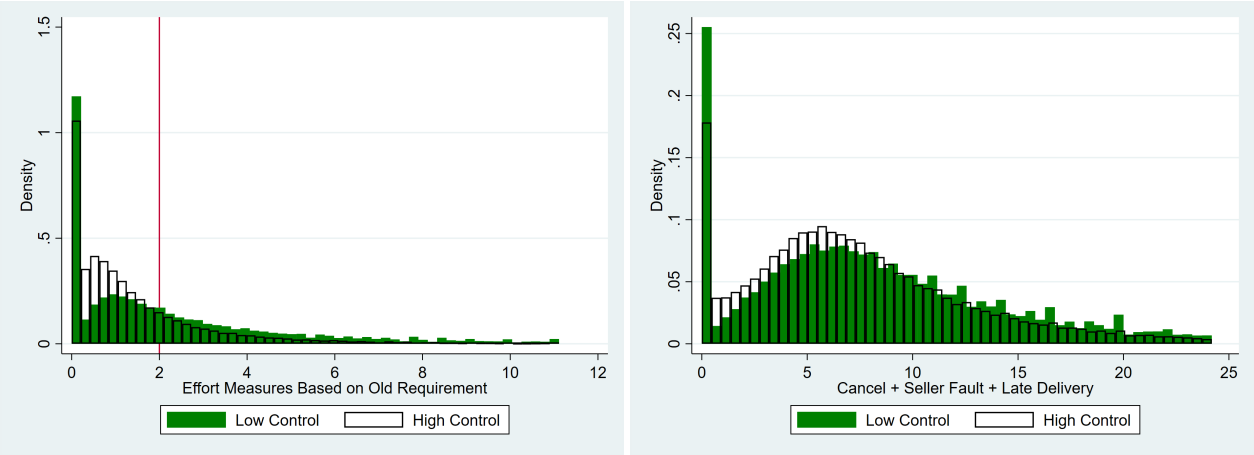
Appendix C The Old Regime Induces Bimodal Effort

As we argue in Section 4.1, under the old certification regime where consumer reports are used, sellers who intend to be badged need to overshoot on administrative data that they can control to overcome the randomness in consumer reports.

In this section, we elaborate on the behavior incentive induced by the old regime, which is based on both consumer reports and administrative data. Facing overall randomness in markets with low controllability, sellers need to overshoot on administrative data on seller behavior in order to overcome the more critical consumer reports. In the meantime, markets with low controllability may also have a higher variance of the randomness, as in our numerical simulation in Figure 8. In that case, sellers with sufficiently high effort costs may give up on exerting effort. Therefore, if we compare the distribution of seller quality in markets with low versus high controllability before the policy announcement, we should see that it has fatter tails in markets with low controllability.

We test this hypothesis in Figure 10. In the left graph, the effort measure is based on the old certification requirements, and is calculated based on the "before" data. The vertical line at 2% indicates the certification threshold under the old regime. The quality distribution in markets with low controllability (represented by solid bars) indeed has fatter tails than that in markets with high controllability (represented by hollow bars). Note that because we proxy effort by the old requirements, which contain consumer reports, there will be larger (negative) randomness in this proxy in markets with low controllability; therefore, the distribution of true effort could have

Figure 10: Quality Distribution of Sellers in Different Markets



Notes: The effort measure is based on the old certification requirements in the left graph, and based on the new certification requirements in the right graph. Both effort measures are calculated using data from the "before" period. Markets are divided into high-controllability and low-controllability based on a median split. Vertical line indicates the certification threshold under the old regime.

even fatter tails in markets with low controllability. To see this, in the right graph, we use effort measures per the new certification requirements, again using the "before" data.[16] Here we see a clearer pattern of fatter tails in the distribution of seller quality in markets with low controllability. Both graphs are consistent with the behavior incentive induced by the old certification requirements, which are in turn consistent with the shape of the simulation curve in Figure 5.

Appendix D Multitasking

The policy change may induce sellers to shirk on consumer-relevant quality dimensions that are not in the new certification requirements. In the spirit of Holmstrom and Milgrom (1991), we test whether focusing on the certification-highlighted quality metrics comes at the cost of deteriorating overall quality, using the RDD specification in equation 1. The sample for this analysis is at the seller-month level, and the dependent variables are share of transactions with negative or neutral ratings, low DSR, and buyer claims, respectively. We focus on sellers right around the threshold because these sellers exhibit strong threshold targeting behavior as a result of the policy change. Given this, these sellers should have the strongest tendency to shirk on other quality dimensions as they make effort towards the new certification, compared to sellers who already satisfy the new requirements (and thus have less incentive to make additional effort).

[16] We do not draw a vertical line here because there are two thresholds for seller defects and late delivery rate.

Table 8 reports the regression results. All dependent variables are in percentage points. We do not find statistically significant coefficient estimates for either NotBadgedSim*$Post_{Announce}$ or NotBadgedSim*$Post_{Implement}$ across all outcome variables. In addition, the size of the estimates is negligible. Therefore, the evidence suggests that sellers do not perform worse in these quality metrics as a result of the policy change, which emphasizes seller inputs. However, these consumer-reported quality measures can be positively correlated with administrative data on seller behavior, which can compensate some of the negative effects due to multitasking. Therefore, we interpret the findings here as no evidence of excessive multitasking or shirking, while noting that our findings can be compatible with some amount of multitasking.

Table 8: Multitasking

	(1)	(2)	(3)
	Neg./Neutral rating	Low DSRs	Buyer claims
NotBadgedSim*$Post_{Announce}$	0.012	0.034	-0.005
	(0.016)	(0.041)	(0.024)
NotBadgedSim*$Post_{Implement}$	-0.003	-0.004	-0.010
	(0.015)	(0.037)	(0.024)
NotBadgedSim*Pre3	-0.014	-0.032	-0.011
	(0.016)	(0.039)	(0.020)
NotBadgedSim*Pre2	0.007	-0.013	0.014
	(0.015)	(0.039)	(0.020)
NotBadgedSim*Pre1	-0.005	0.032	-0.003
	(0.018)	(0.043)	(0.023)
Observations	1,013,079	1,013,079	1,013,079
R-squared	0.085	0.084	0.101
Seller Fixed Effect	✓	✓	✓
Month Fixed Effect	✓	✓	✓

Notes: One observation is a seller-month pair. Outcome variables are in percentage points. Moreover, NotBadgedSim equals 1 if the seller does not meet the new certification requirements on the policy announcement date; $Post_{Announce}$ and $Post_{Implement}$ are the dummies for the months after the policy announcement month and implementation month, respectively; Pre3 is the dummy for Month -13, -12, and -11; Pre2 is the dummy for Month -10, -9, and -8; Pre1 is the dummy for Month -6 and -7. The omitted group in the regression is Month -16, -15, and -14. We also control for the constant term in the regression. Standard errors in parentheses and clustered at the seller level. *** p<0.01, ** p<0.05, * p<0.1.

Appendix E More Results on Certification Premium

In this section, we provide more details on the analyses of certification premium. First, we re-run equation 2 separately for the four market types defined by the median split of controllability and relevance. As shown in Table 9, the two markets of low relevance (i.e., low-relevance-low-controllability and low-relevance-high-controllability) see a similar increase in eTRS premium, of 20–22%. The Wald test of the difference in these percentage effects suggests that these changes are not statistically significant. The two markets of high relevance (i.e., high-relevance-low-controllability and high-relevance-high-controllability), however, experience a much greater increase in eTRS premium (79% and 96%, respectively). Although the Wald test of the difference in these percentage effects suggests that the difference between these changes is not statistically significant, each of these effects is significantly higher than the percentage effects in the low-relevance markets. Therefore, these patterns suggest that consumers' response to the policy change in terms of eTRS premium is mostly pronounced along the margin of relevance, instead of controllability.

Table 9: Certification Premium — Relevance by Controllability

Outcome variable: Dummy for whether a listing results in sales				
	(1) low relevance low controllability	(2) low relevance high controllability	(3) high relevance low controllability	(4) high relevance high controllability
badge	0.060***	0.041***	0.029***	0.016***
	(0.002)	(0.002)	(0.001)	(0.002)
badge * post	0.012***	0.009***	0.023***	0.015***
	(0.002)	(0.003)	(0.001)	(0.002)
Constant	0.160***	0.160***	0.161***	0.159***
	(0.001)	(0.001)	(0.000)	(0.001)
Percentage change	20%	22%	79%	96%
Observations	392,508	717,966	510,776	1,765,911
R-squared	0.138	0.076	0.105	0.074
Matched Set Fixed Effect	YES	YES	YES	YES

Notes: One observation is a listing. The outcome is whether a listing sells; $Post_{Implement}$ are the dummies for the months after the policy implementation month. Listings are matched based on seller ID, listing title and subtitle, Product ID, listing date, and listing price. Standard errors are clustered at the matched listing level. *** p<0.01, ** p<0.05, * p<0.1.

Second, we report the analysis on the price premium of the eTRS badge, where we run a modified version of equation 2 with the logged sales price as the dependent variable. This allows us to identify the eTRS premium in terms of price difference that reflects consumers' willingness to pay for otherwise similar items that differ in the eTRS badge. Specifically, the listings are matched

Table 10: Certification Premium — Consumer Willingness to Pay for eTRS Badge

Dependent Variable	*log(Price)*			*log(Price+Shipping)*		
	(1) all	(2) low relevance	(3) high relevance	(4) all	(5) low relevance	(6) high relevance
badge	0.007***	0.015***	0.002	0.007***	0.014***	0.002
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)
badge*post	0.010**	0.002	0.016***	0.010**	0.002	0.015***
	(0.005)	(0.007)	(0.005)	(0.005)	(0.007)	(0.005)
Percentage change	140%	10%	779%	142%	15%	728%
Observations	26,079	11,426	14,653	26,079	11,426	14,653
R-squared	0.995	0.997	0.993	0.996	0.997	0.993
Matched Set Fixed Effect	YES	YES	YES	YES	YES	YES

Notes: One observation is a listing. The outcome is logged sales price or logged sale price with shipping fee; $Post_{Implement}$ are the dummies for the months after the policy implementation month. Listings are matched based on seller ID, listing title and subtitle, Product ID, and sales date. Standard errors are clustered at the matched listing level. *** p<0.01, ** p<0.05, * p<0.1.

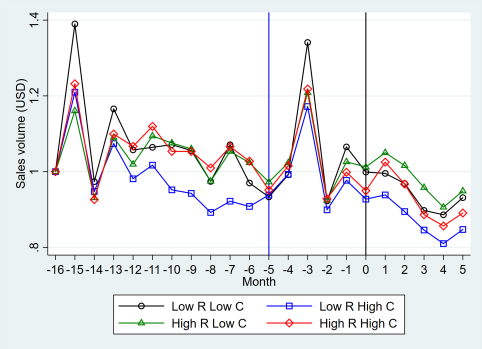
based on seller ID, listing title and subtitle, Product ID, and sales date.

Regression results are reported in Table 10. First, we use the logged sales price as the dependent variable and find that there is an overall positive consumer willingness to pay for the eTRS badge under the old regime, estimated at 0.7% (column 1). The eTRS premium increases by 1 pp after the policy implementation, possibly because of the policy change or time trends. Similar to the analysis of the eTRS premium in the main text, we compare how the eTRS premium changes in low-relevance markets vs. high-relevance markets, to eliminate the assumed similar time trends across markets on eBay. Columns 2 and 3 suggest that the increase in eTRS premium is more pronounced in high-relevance markets than in low-relevance markets. However, the difference between these percentages is not statistically significant, possibly because of the small sample size. In columns 4, 5, and 6, we repeat these analyses with the outcome variable being the logged total price, where the total price is the sales price plus the shipping charge. Similarly, we find that the percent increase in eTRS premium is higher in high-relevance markets than in low-relevance markets (728% vs. 15%, although the difference is not statistically significant). Therefore, the findings here are qualitatively similar to the findings in the main text with sales probability as the outcome variable.

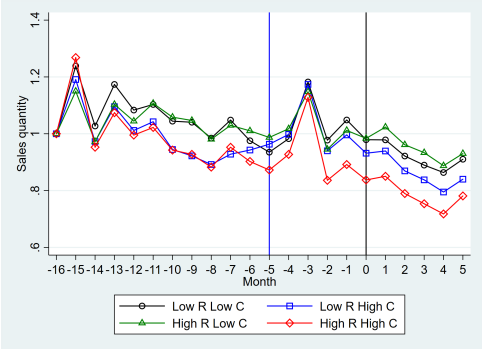
Appendix F Market-level Time Series

In Figure 11, we plot the times series of different outcome variables for the four types based on the median splits of relevance and controllability. The values are normalized by the value in the first month of our sample.

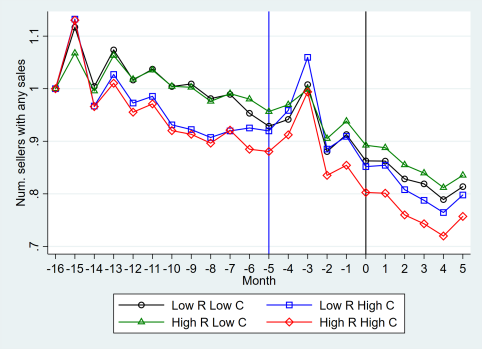
Figure 11: Normalized Time Series of Dependent Variables in Market-level Analysis



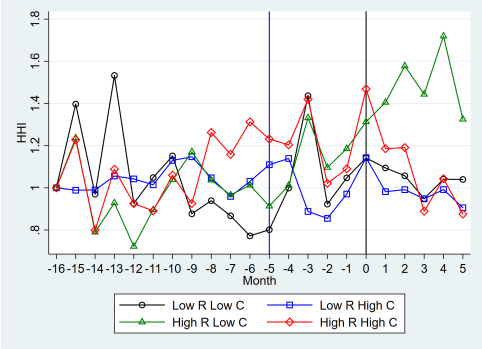
(a) Sales volume (USD)



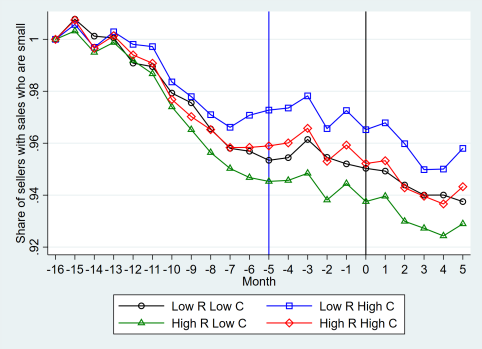
(b) Sales quantity



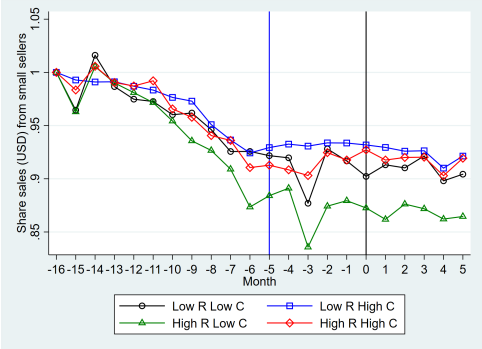
(c) Num. sellers with any sales



(d) Herfindahl–Hirschman index (HHI)



(e) Share of sellers with sales who are small



(f) Share sales (USD) from small sellers

Notes: All variables are normalized by the value in the first month of our sample. Markets are divided into four types based on the median splits of relevance (R) and controllability (C). Blue and black vertical lines indicate the policy announcement and implementation months, respectively.