# Designing Quality Certificates: Insights from eBay *

**Xiang Hui**[†]
Washington University

**Ginger Zhe Jin** [‡]
University of Maryland
and NBER

**Meng Liu** [§]
Washington University

February 22, 2024

**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 consumer reports can be driven by factors not entirely within sellers' control. Alternatively, using administrative data makes the certification more controllable by sellers, but these data track only a subset of seller behavior and may not be fully aligned with consumer experience. To answer the above 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, as well as allowing 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 are more correlated with consumer reports. Lastly, the proportion of certified sellers becomes more homogenized across markets, and sales seem to 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, Alexander MacKay, 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 based on past behavior (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, U.S. News ranking of colleges, and various 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' 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, are automatically collected for record keeping by market regulators (e.g., an e-commerce platform), and tend to focus on seller behavior. Theoretically, there is a relevance–controllability trade-off between these two types of information: While using consumer reports makes the certification more relevant for consumer experience, they may be based on factors not entirely within sellers' control, which could discourage seller effort. The use of administrative data, on the other hand, increases sellers' controllability in managing the certification, but these data track only a subset of seller behavior and therefore may not be fully aligned with consumer experience, and may induce gaming behavior. Therefore, the question we seek to answer is: 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 this question by leveraging rich data from eBay, which is an ideal context for several reasons. First, the eBay Top Rated Seller (eTRS) quality certification is a crucial trust mechanism on the platform, and is the only reputational signal displayed on the search result pages. During our period of study, eBay replaced consumer ratings with administrative data in the criteria for awarding the eTRS certification. This major

redesign creates a rare exogenous shock to the quality certification's information basis in this large e-commerce marketplace. Second, the diverse range of product markets on eBay, each differentially affected by the uniform regime shift, offers insights into the important trade-off between relevance (the usefulness of administrative data to buyers) and controllability (the influence sellers have over consumer feedback and certification status). Lastly, our findings from eBay may be applicable to other firms that also depend on the effort of "agents"—be they employees in a traditional firm, third-party sellers on a digital platform, or arm-length contract partners that supply inputs to the firm. In these different settings, the firm that defines the rating system faces a similar trade-off between measuring agent performance by consumer rating or administrative data.

We start by constructing proxies for relevance and controllability. We conceptualize consumer ratings as the ground truth about the consumer experience of a transaction. The old certification relies on consumer ratings, and the new certification relies on administrative data on seller behavior. The relevance of the new certification refers to its *alignment* with the ground truth. It is measured as the correlation between a seller's certification status (under the old regime) and the counterfactual certification status. This counterfactual status is determined by applying the new requirements to the seller's past behavior as of the policy announcement date. This relevance metric corresponds to the Pearson correlation (also known as the Phi coefficient) in a two-by-two confusion matrix in the machine learning literature. Intuitively, in a high-relevance market, certification is awarded to a similar set of sellers under both certification regimes, meaning the new certification aligns well with the consumer experience in that market.

Sellers' controllability in managing their certification status concerns the *variability* of the mapping between seller effort and consumer experience. This variability arises because consumer experience can be influenced by factors not entirely within sellers' control.[1] Specifically, we measure the (lack of) controllability as the variance of a seller's certification status (under the old regime) for a given level of seller effort. This effort is measured using the metrics from the new regime, which eBay deemed critical.[2] Greater variance implies less controllability. Intuitively, a low-controllability market is one where sellers face a larger variability in getting certified (owing to receiving negative consumer ratings) conditional on seller effort.

[1] An example is logistic delay due to weather: consumers may leave negative ratings because they did not receive their items on time, even if the delay is due to unforeseen weather conditions and the seller had handled the order promptly.

[2] The ideal case would be to use all dimensions of seller effort, which we do not observe. Our assumption is that a market's ordinal ranking of controllability is the same regardless of whether it is defined by the dimensions that are deemed critical by eBay or by all dimensions of seller effort.

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—a type of gaming behavior based on Baker (1992). On the buyer's 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 proportion of certified sellers is more homogenized across markets. But there is little change in the total number of transactions, perhaps because the increase in transactions owing to higher seller effort offsets the decrease in transactions due to the lower relevance of the new certification in markets with lower controllability. However, in markets with lower controllability, sales seem to become more concentrated towards larger sellers, likely because of eBay's institutional details discussed in Section "Homogenization of Share of Certified Sellers across Markets".

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 could be ambiguous because of the threshold effect and, more generally, potential gaming behaviors from sellers. Third, whether consumers find the quality certification useful depends on its alignment with consumer reports. Lastly, the design of quality certification could 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 that paper was published, more authors have used natural experiments to study the effect of information disclosure on market outcomes, such as calorie posting in chain restaurants (Bollinger et al., 2011), labeling on food products (Bai, 2018), occupational licensing on a home service platform (Farronato et al., 2020), and licensing of food sellers on Alibaba (Jin et al., 2022). In the eBay context, previous work has shown that consumers value the eTRS badge (Elfenbein et al., 2015; Hui et al., 2016), and the design of the badge, both the stringency and granularity of the certification threshold, can have

large effects on the entry of sellers and their quality provision on eBay (Hui et al., 2017, 2020). Our paper contributes to this literature by studying another key design question, namely, the trade-off of using different information sources in constructing quality certificates.

Next, our paper is also related to a large literature that studies consumer reports and reputation systems. The literature finds that reputation systems lead to higher demand and discipline seller behavior in various contexts. Examples include e-commerce (Chevalier and Mayzlin, 2006; Fan et al., 2016; 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), and eBay (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) (see Dellarocas (2003) and Tadelis (2016) for summaries). Our paper contributes to this literature by showing the pros and cons of using consumer reports, and suggests that market designers need to weight the informational benefit of consumer reports against its negative effect on seller effort when constructing trust signals. These results also 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).

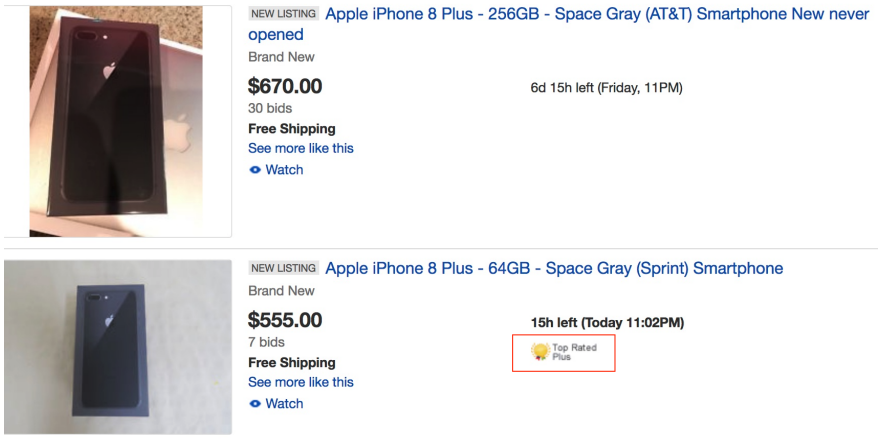
Lastly, our paper relates to the central topic in the economics literature on incentive contracts and principal-agent problems. At its core, there is incentive misalignment between the two parties, and the principal needs to design an incentive scheme to motivate the agent (Holmström, 1979; Baker, 1992; Holmstrom and Milgrom, 1991). In this literature, the most related paper to ours is Baker et al. (1994), who studies the use of subjective performance measures and objective measures in incentive contracts. The trade-off in their paper is similar to our empirical findings on the trade-off between using consumer ratings and administrative data. We contribute to this literature by applying the theory to a novel and important context, i.e., the design of quality certification, which is ubiquitously used in everyday life (Dranove and Jin, 2010).

## 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 a high level, eBay aggregates a number of seller performance metrics to determine whether the seller meets the threshold for the eTRS certificate. We discuss the breakdown of the

Figure 1: Example: Search Results Page




metrics in Table 1.

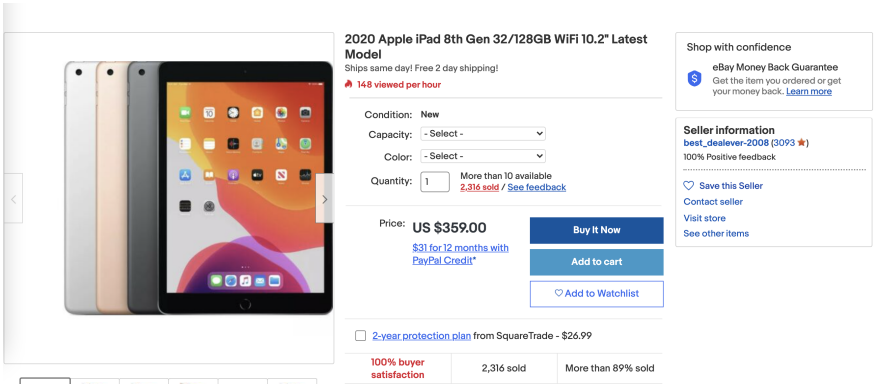
The main benefit of having an eTRS certificate is that sellers have the eTRS badge displayed on their listings on the search results page (e.g., Figure 1) and on the listing page (shown after clicking on the item on the search results page). This benefit is valuable because the eTRS certificate is the only reputation signal shown on the search results page, and it 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 to the listing page (e.g., Figure 2), where consumers can see a seller's percentage positive ratings. On the listing page, an additional click on the seller profile name leads to the seller profile page, where consumers can see a seller's Detailed Seller Ratings (DSR) 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, according to Nosko and Tadelis (2015), who used clickstream data from eBay, less than 1% of users ever go to the seller's profile page, which implies that DSR are not very visible to most users.

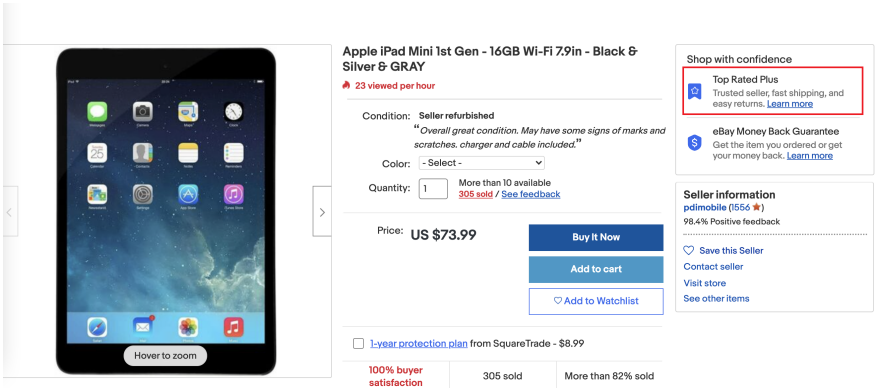
In 2015, eBay redesigned the eTRS program, aiming to create a simpler and more objective standard. The new eTRS criteria were announced in September 2015 and implemented in February 2016, and the changes are summarized in Table 1.[3] At a high level, consumer reports (metrics 1, 2, 3, 5) were replaced with administrative data that emphasize seller behavior. More specifically,

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

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

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. In both regimes, the requirements are categorized into two groups: defect (non-logistics-related metrics) and late delivery (logistics-related metrics). 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 certified sellers after the policy change. The requirement on minimum past sales is unchanged: \$1,000 in value and 100 transactions in the past evaluation period, which is the past three months if a seller has at least 400 transactions in that period, or past 12 months if otherwise.

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')[4]; and late delivery is now determined using tracking information provided by post offices instead of consumer reports (metric 5' instead of 5). Specifically, an item is classified as late if the duration between the post office's receipt of the item and the buyer's payment exceeds the seller-specified handling time. As a result of the regime change, sellers should find it easier to manage and predict their eTRS status. For example, a seller eager to earn the eTRS certification may offer refunds in case of claims (which would be recorded as resolved claims by eBay), refrain from cancellation, and send the items to the post office within the specified handling time.

Note that even after the policy change, the percentage of positive ratings (metric 1) and DSR (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

Our research question is: What would be the effect of using consumer ratings vs. administrative data in constructing quality certification? In this section, we provide a conceptual framework to

[4]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.

better characterize the effect of the regime change.

We conceptualize consumer ratings as the ground truth about the consumer experience of a transaction. In doing so, we abstract away from imperfections in using consumer ratings to measure consumers' true experience.[5] Despite these imperfections, consumer ratings are the *best available* representation of consumer experience, i.e., they are closer to the true, unobserved consumer experience than other alternative quality measures, such as measures of seller effort. Our qualitative insights regarding the treatment effect heterogeneity along the dimensions of relevance and controllability dimensions are likely to be applicable in markets where consumer ratings more accurately reflect the true, unobserved consumer experience compared to other alternative quality measures, such as seller effort. These insights may not hold in markets where consumer ratings are poorer representation of consumer experiences than measures of seller behavior.[6]

For clarity of presentation, we assume that the old certification relies entirely on consumer ratings, and the new certification relies entirely on administrative data about a subset of seller behaviors. The quality certification on eBay is binary, i.e., a seller is either certified or not. Given the binary nature of certification, we can characterize the relationship between the two regimes in a confusion matrix. In Table 2, the parameter in each cell represents the probability of that cell in the whole population of sellers: $P$ represents true positives or true negatives (i.e., agreements between the two regimes), and $A$ and $B$ represent false negatives and false positives (i.e., disagreements between the two regimes), respectively.[7] Note that by definition, the four cells sum up to 1, namely, $2P + A + B = 1$, so we have two degrees of freedom.

[5]The imperfections include various types of rating biases (Dellarocas and Wood, 2008; Bolton et al., 2013; Filippas et al., 2022; Moe and Trusov, 2011; Brandes et al., 2019; Hui et al., 2022) and fake reviews (Mayzlin et al., 2014; Luca and Zervas, 2016; He et al., 2020).

[6]We made this assumption also for practical reasons. We have three unknown parameters: the true consumer experience, the variance when using consumer ratings to approximate this true experience, and the variance when using administrative data to approximate this true experience. However, we have only two data sources: consumer ratings and administrative data. Therefore, we need to make identification assumptions. As discussed, we chose to assume that consumer ratings fully reflect the true consumer experience because this assumption is parsimonious and reasonable. Even though consumer ratings are not perfect, they seem to be the most fitting representation of consumer experience available to us (and to the platform). Consequently, the platform in practice uses and targets this proxy.

[7]We assume the two diagonal terms are equal, for simplicity. Having the diagonal terms to be different from each other does not affect the main insights; we assume them to be equal (instead of off-diagonal terms) because these represent "infra-marginal" sellers whose status did not change and are in some sense less interesting to study compared to the ones whose status changes. The key qualitative insights remain similar if we allow diagonal values to differ, as we discuss in Section 1.2 in the appendix.

Table 2: A Confusion Matrix

| | New (counterfactual) badge =1 | New (counterfactual) badge =0 |
|---|---|---|
| Old (actual) badge=1 | $P$ | $B$ |
| Old (actual) badge=0 | $A$ | $P$ |

Consumers' transaction experience has two components: seller effort in providing the best quality of products and services, and how quality provision translates into consumer experience. The latter may depend on factors that are not entirely within sellers' control, e.g., weather and how carefully shippers handle packages during transportation. When these external factors are large (from the seller's perspective), we have larger $A$ and $B$.

The relevance of a certification refers to its alignment with consumer experience. It is measured as the correlation between a seller's certification status under the old regime (which by assumption is based entirely on consumer ratings, i.e., the ground truth of consumer experience) and her counterfactual status under the new regime. Pearson correlation, or Phi coefficient, is the standard way to measure the quality of binary classification in the machine learning literature (Bishop and Nasrabadi, 2006). This measure considers all values in a confusion matrix (i.e., true positives, true negatives, false positives, and false negatives) and gives a balanced representation of the alignment of two classifications.

As shown in the first row of Table 3, the relevance of the old badge for consumer experience is 1 by definition, and the relevance of the new badge is $\frac{P}{(P+A)(P+B)} - 1$ (see derivation in Appendix Section "Deriving Relevance and Controllability"), which ranges from -1 (lowest relevance) to 1 (highest relevance).[8] Based on the expression, the relevance of the new badge is highest, or 1, if it perfectly aligns with the old badge, i.e., $A = B = 0$ and $P = 0.5$. The relevance of the new badge is lowest, or -1, if it never agrees with the old badge, i.e., $P = 0$, $A + B = 1$.

The controllability of a certification refers to the variability of the mapping between seller effort and certification status. It is measured based on the variance of a seller's certification status for a given level of seller effort. Because we cannot observe all dimensions of seller effort, we use the dimensions that eBay deemed critical under the new regime. We assume that a market's ordinal ranking of controllability is the same whether defined by these dimensions or by all dimensions of seller effort.

[8]Our implicit assumption is that consumers lack the sophistication to reverse the meaning of a signal when this signal is consistently wrong, i.e., they cannot interpret the absence of a signal as a new signal, or conversely, to see the signal as the absence of one.

Table 3: Relevance and Controllability in the Two Regimes

| | Old (actual) badge | New (counterfactual) badge |
|---|---|---|
| Relevance | 1 | $\frac{P}{(P+A)(P+B)} - 1$ |
| Uncontrollable Variation | $\frac{1}{2}(\frac{PA}{(P+A)^2} + \frac{PB}{(P+B)^2})$ | 0 |

Controllability is measured as the negative of "uncontrollable variation" from the seller's perspective. Because the new badge is entirely based on seller effort, by definition it has no uncontrollable variation, as shown in the second row of Table 3. In comparison, the old badge has an uncontrollable variation of $\frac{1}{2}(\frac{PA}{(P+A)^2} + \frac{PB}{(P+B)^2})$, which captures the variance in certification status conditional on seller effort (see details in constructing this measure in Section "Data"). Based on the expression, the sellers face the biggest uncontrollable variation (and thus the least controllability) when $P = A = B = 0.25$, meaning that the new certification does no better than random guessing consumer experience (measured by the old certification based on consumer ratings).

As can be seen from Table 3, the theoretical trade-off between using the two regimes is relevance vs. controllability. The column corresponding to the old badge shows that while using consumer ratings to construct quality certification makes it directly relevant for consumer experience conditional on seller effort, doing so may discourage sellers from exerting effort because of the lower controllability of their certification status. The column corresponding to the new badge shows that using administrative data allows sellers to have maximum controllability of their eTRS status, which should incentives their effort along the highlighted dimensions, but it makes the new badge less relevant than the old badge to buyers.

What would happen to sellers, buyers, and market equilibrium after the regime change? From the perspective of sellers, this allows for a more precise prediction of how their efforts will affect their certification status, and consequently, their expected returns on effort. This increased precision motivates sellers below the certification threshold to increase their efforts, while those above to shirk, consistent with classical principal-agent theory. Moreover, seller effort can be multifaceted, with enhancements in certain effort dimensions potentially coming at the expense of others, as highlighted by Holmstrom and Milgrom (1991) in their multitasking theory.

On the buyers' side, if seller effort does not change, buyers may value the new quality certificate less than the old one, and the valuation should decrease more in markets where the administrative data are less relevant to them (i.e., less aligned with their transaction experience). The change in consumers' valuation for the badge may in turn have long-term implications on consumer retention.

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 experience. 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 redesign to shed light on the research question.

# 3 Data

We use proprietary data from eBay, a popular e-commerce marketplace.[9] 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).[10] 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 after 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 the normalized time series of key variables, where the normalization is taken

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

[10] 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.

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 in Appendix Section "Market-level Time Series".

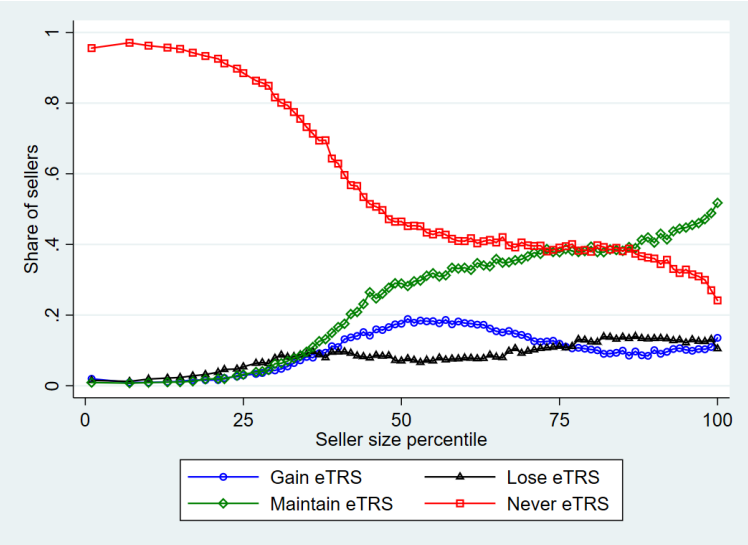
Additionally, in Figure 3 we present the relative changes in the share of sellers who lost, gained, or maintained their badge status because of the policy implementation. To categorize sellers, we examine their monthly badge status during the six months before and after the regime change. Specifically, we assess whether they were badged for more than half of the time in either the before or after period. For instance, a seller badged for more than three months in the before period and also more than three months in the after period is categorized as a "Maintain eTRS" seller. Conversely, a seller badged for less than or equal to three months in the after period is categorized as a "Lose eTRS" seller. The definitions for the other categories follow a similar logic. Note that for each seller size percentile, the shares in the four categories sum up to 1.

Figure 3 reveals several interesting patterns. Smaller sellers are not significantly affected by the regime change relative to larger sellers. For instance, among sellers in the lowest quartile by size, over 90% fall into the "Never eTRS" category. Next, midsize sellers appear to benefit the most from the regime change, as a larger proportion of them acquire the badge compared to other size groups, with the peak of the "Gain eTRS" curve occurring for sellers around the median size. Lastly, larger sellers have a higher proportion able to retain the badge compared to sellers in other size groups, as indicated by the "Maintain eTRS" curve peaking for sellers in the top size quartile.

Because our goal is to understand the trade-off between using different metrics, we first study the correlation between the various metrics based on consumer reports and administrative data. We create a heat map of these metrics, shown in Figure 4, where a darker shade 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 has 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 has a low Detailed Seller Rating on shipping speed from consumers.

Examining the figure reveals a few patterns: First, the consumer reports metrics (i.e., negative feedback, low DSR on item description, buyer claim, and low DSR on shipping speed) are positively correlated with one another. For example, negative feedback is highly correlated with low DSR on item description, suggesting that a major source of customer complaint is unclear item description. Second, the correlations between consumer reports and administrative data on seller behavior

Figure 3: Share of Sellers Who Gain, Lose, Maintain, or Never Get Certification




*Notes*: This figure plots the share of sellers who gained, lost, maintained, and never obtained eTRS badge, by seller size percentile using data from the six months before and after the regime change. The categories are defined based on whether sellers were badged for more than half of the time in either the before or after period. For instance, a seller badged for more than three months in the before period and also more than three months in the after period is categorized as a "Maintain eTRS" seller. The other categories are defined using a similar logic. For each seller size percentile, the shares in the four categories sum up to 1.

(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; 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 between using consumer reports and administrative data for constructing quality certificates, we categorize a large number of product markets into different groups depending on the relevance and controllability measures. The product markets follow the product subcategories defined by eBay. Examples of subcategories are Video Games, Women's Clothing, and Cell Phones & Smartphones. To remove outliers, we exclude from our analysis all markets with less than $300,000 in sales revenue in the first three months of the sample period. The remaining 328 markets account for 99.7% of the total sales revenue on the platform.

We present the categorization of markets by controllability and relevance in Figure 5. Each circle represents a product market, with circle size representing total dollar sales. We construct the

Figure 4: 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.

proxy for controllability using the steps below. Recall that we use the dimensions of seller effort that eBay deems crucial and the associated threshold under the new regime as a proxy for good seller behavior.

1. Conditional on sellers meeting the new requirements in Table 1 (which implies high seller effort by definition), use the dummy "*eTRS_goodInput*" to denote whether a seller has an eTRS certificate under the old system.

2. Compute the variance of "*eTRS_goodInput*" across sellers of a given market, weighted by the number of transactions of the seller in that market in the three months before the policy announcement.

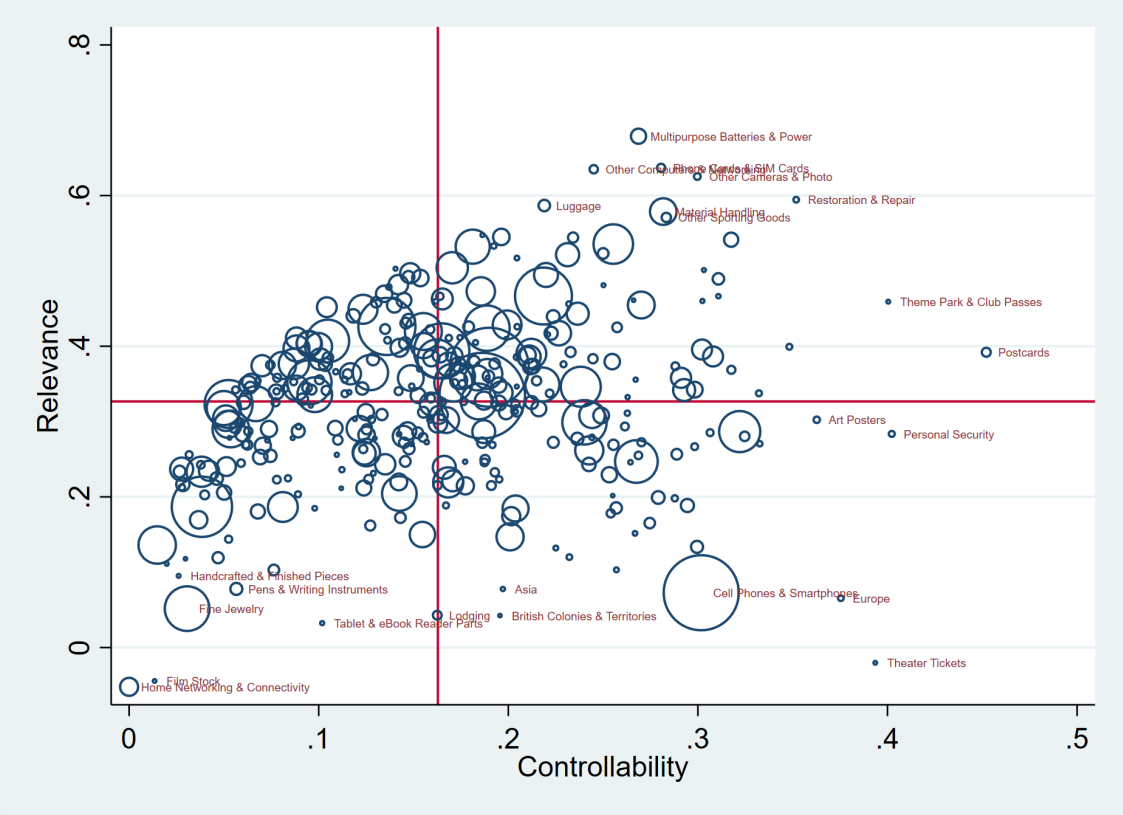
3. Repeat steps 1 and 2 for "*NoneTRS_badInput*" conditional on sellers not meeting the new requirements (which implies low seller effort by definition), where "*NoneTRS_badInput*" denotes whether a seller was not badged under the old system. Then compute the weighted variance as in Step 2.

4. Average over the two variances to construct the variable "*uncontrollable variation*" at the market level.

5. Define controllability for market k as [$max$(*uncontrollable variation*) − *uncontrollable variation* of market k], where the $max$ operation is taken over all markets. Greater controllability corresponds to a smaller uncontrollable variation in a market.

Intuitively, a low controllability in a market means a larger variance of getting the eTRS badge in the old regime conditional on a given level of seller behavior.

Figure 5: Markets Differ in Controllability and Relevance




*Notes*: Relevance is the correlation between a seller's actual badge status under the old regime and her counterfactual badge status by applying the new requirements on her performance as of the policy announcement date. Controllability is the negative of uncontrollable variation, defined as the variance of a seller's badge status under the old regime, conditional on a given level of seller effort; this effort is measured based on whether the seller meets the new requirements. Circle size represents total dollar sales of a market in the three months before the policy announcement. The two lines represent median splits of the two measures.

To construct the proxy for relevance in a market, we look at all sellers in that market and calculate the Pearson correlation coefficient between each seller's certification status (under the old regime) and her counterfactual certification status had the policy change happened immediately at the announcement. The counterfactual badge status is constructed by applying the new requirements to seller performance immediately at the policy announcement so that the sellers have not 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 aligns well with consumer reports, as required in the old certification in that market. The two measures, relevance and controllability, have a correlation of 0.28.

Based on the median split of the two measures (the two lines in Figure 5), we group markets into four types. For example, the Other Computers & Networking and Phone Cards & SIM Cards categories rank high on relevance, and the Handcrafted & Finished Pieces and Fine Jewelry categories rank low on controllability. To understand the characteristics of markets that differ in relevance and controllability, we regress these two measures 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 4. 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 containing highly standardized products (e.g., Prepaid Gaming Cards) score high on relevance. Column (2) shows that 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 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)). Additionally, larger package size also

Table 4: Predictors of Controllability and Relevance

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | relevance (continuous) | relevance (continuous) | relevance (continuous) | control (continuous) | control (continuous) | control (continuous) |
| log(price) | -0.002 | -0.009 | -0.011 | -0.028*** | -0.007 | -0.014** |
| | (0.008) | (0.008) | (0.009) | (0.006) | (0.006) | (0.006) |
| standardized | 0.040*** | 0.039*** | 0.038** | -0.010 | -0.009 | 0.002 |
| | (0.015) | (0.015) | (0.016) | (0.012) | (0.011) | (0.012) |
| log(package size) | | 0.014* | 0.020*** | | -0.041*** | -0.036*** |
| | | (0.007) | (0.007) | | (0.005) | (0.006) |
| log(HHI) | | | 0.020*** | | | 0.016*** |
| | | | (0.007) | | | (0.005) |
| log(total quantity) | | | 0.012** | | | -0.001 |
| | | | (0.006) | | | (0.004) |
| Constant | 0.307*** | 0.254*** | -0.008 | 0.285*** | 0.439*** | 0.381*** |
| | (0.031) | (0.041) | (0.109) | (0.025) | (0.031) | (0.081) |
| Observations | 325 | 325 | 325 | 325 | 325 | 325 |
| R-squared | 0.022 | 0.033 | 0.062 | 0.063 | 0.202 | 0.241 |

*Notes*: One observation is a product category. Outcome variables are (continuous) relevance and controllability of a market. 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.

correlates with lower controllability, possibly because of a larger variance in delivery time and hence consumer ratings. The fact that these regression results are intuitive increases our confidence in market categorization.

# 4 Results on Sellers

Recall from Section "Conceptual Framework" that various forces shape seller behavior. The new certification regime, by reducing the variability in how seller effort translates to certification status, should theoretically encourage those below the certification threshold to increase their effort, and those above it to shirk, according to classical principal-agent theory. Moreover, sellers may focus on improving aspects of their performance that directly align with the certification criteria, potentially at the expense of others. Consequently, the overall impact of the regime change on seller effort remains ambiguous. In this section, we conduct an empirical analysis to identify any changes in seller behavior.

### 4.1 Selection Effect and Changes in Behavior

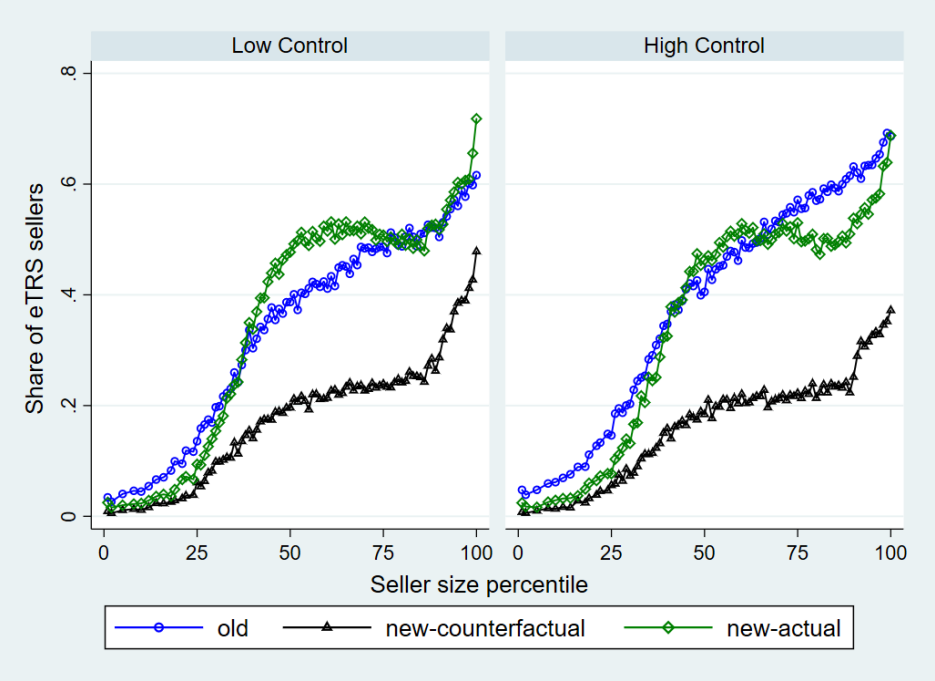
The overall effects of the regime change on sellers can be decomposed into an immediate selection effect (i.e., changes in the composition of certified sellers before any changes in seller behavior) and seller behavior changes in response to the new policy. To measure the selection effect, we compute a seller's counterfactual 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 effort in response to the new policy, so the difference between the seller's actual and counterfactual eTRS status should capture only selection.[11] To measure the changes in seller behavior in response to the new policy, we focus on the difference between a seller's counterfactual eTRS status on the policy announcement date and the eTRS status on the policy implementation date. For example, if a seller is not eTRS certified by the counterfactual computation but later becomes eTRS certified, we infer that the seller has exerted effort based on the eTRS requirements between the policy announcement and implementation dates.

In Figure 6, 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 Figure 2 and the insights are similar. In the figure, we plot the "old" (Month -5), "new-counterfactual" (at Month -5),[12] and "new-actual" (Month 5) shares of certified 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 for two reasons: First, the variance of the mapping between seller effort and certification status differentially affects sellers with different sizes, per the law of large numbers. Second, 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 badge. In Appendix Section "Seller Size" we elaborate on why seller size is important for our analysis.

[11] Note that there could be measurement errors when we compute a seller's counterfactual 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 was 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.

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

Figure 6: Policy Effect on Sellers by Controllability




*Notes*: The two markets are defined based on a median split on the seller controllability measure. Counterfactual share of certified sellers (represented by triangles) is normalized by a constant; "old" refers to a seller's certification status in Month -5; counterfactual computation is done, giving a seller's counterfactual certification status in Month -5; "new-actual" refers to a seller's certification status in Month 5.

Figure 6 shows that the selection effect is generally negative, because the counterfactual 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 in low-controllability markets than in high-controllability markets, as the negative gap between the counterfactual curve and the "old" curve is smaller in low-controllability markets (on average -15.5%) than in high-controllability markets (on average -22.3%), which holds true for all seller sizes. In Appendix Figure 2 we provide a detailed discussion of market type, seller size, and the selection effect. As can be seen from the graph, the smaller negative gap is jointly explained by a lower "old" curve (on average 32.0% vs. 38.5%) and a higher counterfactual curve (on average 16.5% vs. 16.2%) in low-controllability markets vs. high-controllability markets: The reason for a lower "old" 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 counterfactual 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 Section "The Old Regime Induces Bimodal Effort").

Next, we infer changes in seller behavior from the difference between the "new-actual" and counterfactual shares of certified sellers. From the figure, we see that overall, sellers 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 certified sellers after the policy implementation in these markets; graphically, the difference is 17.2% and 18.7%, respectively, for low-controllability and high-controllability markets. This pattern suggests that overall quality improvement is less pronounced among sellers in low-controllability markets, suggesting that the reduction in effort from sellers above the threshold outweighs the increased effort from those below it. Note that this threshold targeting behavior can be seen as a type of gaming behavior using George Baker's definition: "take actions that increase payouts from the incentive contract without improving actual performance" (Baker, 1992), and this behavior has been documented in other contexts, e.g., (Hunter, 2020; Alé-Chilet and Moshary, 2022).

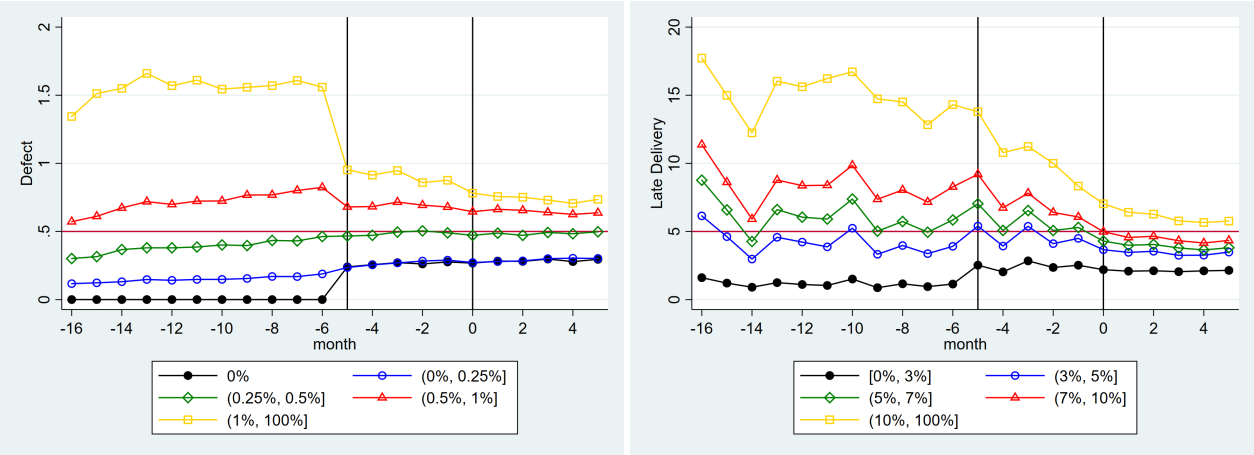
Given that the two curves for the after period look similar across the low-controllability and high-controllability markets, we hypothesize that the smaller quality improvement in low-controllability markets is a result of the threshold effect. We test this hypothesis in the next section.

### 4.2 Threshold Effect

In Figure 7 we plot the time series of seller effort for different seller types. The two effort measures correspond to "defect" and "late delivery" under the new regime described in Table 1, with the horizontal lines denoting the respective thresholds. Seller types are defined based on a seller's performance on either defect or late delivery in the period before the policy announcement. The vertical lines in the future denote the policy announcement month and implementation month, respectively. In both graphs, we see a convergence towards the new thresholds for all seller types. Specifically, in the left graph, sellers with a defect rate lower than the certification threshold before the policy announcement——namely, groups "0%", "(0%, 0.25%]", and "(0.25%, 0.5%]" groups—tend to shirk after the announcement. Conversely, sellers with a defect rate higher than the certification threshold——namely, the "(0.5%, 1%]" and "(1%, 100%]" groups—improve their effort after the announcement. A similar pattern is observed in the right graph, where sellers who previously excelled in fast delivery reduce their effort, while those with a late delivery rate above the certification

threshold improve their shipping performance after the policy announcement. We quantify the changes in behavior across different seller groups in Table 12 in the appendix, and the regression results are consistent with the insights from the graph.

Figure 7: Threshold Effect: Sellers' Effort Adjustment Towards Certification Threshold




(a) Defect Rates of Sellers by Their Performance in the "Before" Period

(b) Late Delivery Rates of Sellers by Their Performance in the "Before" Period

*Notes*: The two effort measures correspond to the two sets of requirements under the new regime ("defect" and "late delivery" in Table 1), with the horizontal lines indicating the thresholds. Each curve represents a seller group based on the sellers' performance on the same measures in the before period. The vertical lines correspond to the policy announcement month (left) and implementation month (right).

To achieve a cleaner identification of the threshold effect, we employ a regression-discontinuity design (RDD), comparing sellers with metrics near the certification threshold, thereby ensuring their underlying qualities are comparably similar along the focal dimensions. Considering the new certification requirements are categorized into defect and late delivery, we construct our estimation sample by identifying sellers who meet the requirements on defect rate but have a late delivery rate between 4% and 6% (i.e., within 1% around the 5% threshold) based on data in the before period. We focus on sellers who meet the defect rate requirement but vary in meeting the late delivery requirement, because the latter is notably more binding: Specifically, there are significantly fewer sellers who meet the stringent late delivery rate but not the defect rate requirement, making the

analogous RDD test less powerful. To estimate the threshold effect, we use the following regression:

$$\begin{aligned} Y_{it} = &\gamma_1 \times Post_{Announce,t} \times NotBadgedCounterfact_i \\ &+ \gamma_2 \times Post_{Implement,t} \times NotBadgedCounterfact_i \\ &+ \sum_{q=1}^{q=3} \beta_q \times Pre_{q,t} \times NotBadgedCounterfact_i + \eta_i + \xi_t + \epsilon_{it}, \end{aligned} \tag{1}$$

where $Y_{it}$ is the outcome variable for seller $i$ in month $t$; $NotBadgedCounterfact_i$ equals 1 if seller $i$ does not meet the new certification requirements on the policy announcement date based on the counterfactual computation; $Post_{Announce,t}$ 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; $\eta_i$ and $\xi_t$ are seller and month fixed effects, respectively. We divide the 11 months in the before period into four periods to maintain a quarter-long duration for each, with the exception of the final one: $Pre_{3,t}$ 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; Months -16, -15, and -14 are the omitted baseline.

Our parameters of interest are $\gamma_1$ and $\gamma_2$. They capture the difference in the temporal changes in effort between sellers that were immediately selected for the eTRS certification by the new regime upon the announcement of the new policy and the sellers that were not automatically qualified for the eTRS certification at the policy announcement time. The $\beta$s capture any preexisting 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 estimated coefficients should be interpreted relative to the difference in the baseline outcome in these three months.

Results are reported in Table 5. Column (1) shows that sellers who are counterfactually not badged (i.e., those with a preexisting average late delivery rate between 5% and 6%) improve on delivery speed relative to sellers who are counterfactually badged (i.e., those with a preexisting average late delivery rate 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 $Post_{Implement}$, and the coefficient measures the additional effect on top of the coefficient on $Post_{Announce,t}$. 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.

Table 5: Threshold Effect

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| | Threshold Effect | | | Placebo Effect | | |
| | late delivery | late delivery | late delivery | defect | defect | defect |
| NotBadgedCounterfact×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) |
| NotBadgedCounterfact×Post$_{Implement}$ | | -0.313*** | -0.313*** | | -0.006 | -0.006 |
| | | (0.058) | (0.058) | | (0.023) | (0.023) |
| NotBadgedCounterfact×Pre3 | | | 0.047 | | | 0.002 |
| | | | (0.067) | | | (0.007) |
| NotBadgedCounterfact×Pre2 | | | -0.015 | | | 0.016** |
| | | | (0.071) | | | (0.008) |
| NotBadgedCounterfact×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 NotBadgedCounterfact 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.

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 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.

We also conducted the alternative RDD analysis on the group of sellers who meet the late delivery requirements but have a defect rate between 0.4% and 0.6%. The estimates yield qualitatively the same results but are not statistically significant, because the sample size is seven times smaller. These findings are reported in Appendix Table 3.

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 Section "Multitasking").

# 5 Results on Buyers

As discussed in Section "Conceptual Framework", from the buyers' perspective, conditional on seller effort, buyers may value the new quality certificate less than the old one, especially in markets where the administrative data are less relevant to consumers, i.e., less reflective of their transaction experiences. This could potentially deter future purchases on the platform. Conversely, if consumers incorporate policy-induced changes in effort in equilibrium, they might value the new quality certificate more, despite its reduced relevance, potentially increasing their likelihood of returning to the platform to make a purchase. Given the ambiguous theoretical predictions, we proceed to examine the empirical evidence in this section.

## 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 are 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[13], 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 the sales probability and 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 have the *eTRS badge* visible 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

[13]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.

experiments, where sellers randomize the appearance of the 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:

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

where $Solds_{ij(t)}$ 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.

We start by plotting the time series of the eTRS premium, which is the difference in the average probability of a sale between badged and non-badged listings, for low-relevance and high-relevance markets. As can be seen from Figure 4 in the appendix, although the two series are not perfectly parallel, there is a clear increase in the eTRS premium in high-relevance markets compared to low-relevance ones.

Next, we report the regression results in Table 6. The first column shows that consumers are 3.5 percentage points more likely to purchase a listing with the eTRS badge than 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 increases 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 either high or low relevance, we can compare consumers' valuation for the quality certificate across markets. For 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 value in quality certification. The same analyses of certification premium by the four market types defined by relevance-by-controllability can be found in Appendix Section "More Results on Certification Premium".

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 4, low-relevance markets tend to be those

Table 6: Certification Premium

| *Outcome variable: Sold* | | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| | all markets | low relevance | high relevance |
| badge | 0.035*** | 0.053*** | 0.026*** |
| | (0.001) | (0.001) | (0.001) |
| badge×Post$_{Implement}$ | 0.016*** | 0.010*** | 0.020*** |
| | (0.001) | (0.002) | (0.001) |
| Percentage change ($\Delta$) | 46% | 20% | 78% |
| $H_0 : \Delta_{low\ rel} = \Delta_{high\ rel}$ | | | |
| Wald test (p-value) | | p<0.001 | |
| 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. The p-value of the Wald test on the difference in the percentage increase in low- vs. high- relevance markets is reported. Standard errors are clustered at the matched listing level. *** p<0.01, ** p<0.05, * p<0.1.

with less standardized products, and consumers should value ratings (and thus quality certificates) more in these markets because consumers cannot easily assess quality using 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 are more aligned with consumer reports.

A concern for the identification of the badge premium is changing consumer preferences for fast shipping and returns between non-holiday and holiday seasons. To mitigate this concern, we estimate the badge premium across different months for both low- and high-relevance markets. Reassuringly, we observe a relatively stable badge premium in November, December, and January during the pre-period, suggesting that changes in consumer preference are not dramatic. In comparison, there was a noticeable increase in the badge premium following policy implementation in high-relevance markets, relative to low-relevance markets, consistent with the key insight from Table 6.

As a robustness check, we rerun Equation 2 to identify an alternative definition of the eTRS

premium (i.e., the difference in equilibrium sales price for otherwise-similar transactions that differ only in the eTRS badge). This 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 November 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 Section "More Results on Certification Premium".

### 5.2 Consumer Retention

Having studied the change in consumers' valuation for the badge, we now study the downstream consequence of using administrative-data-based quality certification in terms of consumer retention. We run the following regression analysis on the sample at the market-month level:

$$Y_{mt} = \beta_1 \times Relevance_m \times Post_{Implement,t} + \beta_2 \times Controllability_m \times Post_{Implement,t} + \eta_{m,\tilde{t}} + \xi_t + \epsilon_{mt}, \tag{3}$$

where $Y_{mt}$ is the average buyer retention rate (whether a buyer purchases another item on eBay, regardless of product category) in market $m$ in month $t$; $Relevance_m$ and $Controllability_m$ are dummies for whether market $m$ has above-median relevance or controllability; $\xi_t$ 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; and $\epsilon_{mt}$ are idiosyncratic errors.

We first plot the time series of the average consumer retention rate (multiplied by 100) for high-relevance and low-relevance markets in Figure 5 in the appendix. The graph shows strong evidence of parallel pre-trends of the retention rate between the two groups, and a narrowing gap after policy implementation.

We then report the regression results in Table 7. Column 1 shows that in the markets where

input measures are more relevant for consumer experience, 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 are more aligned with consumer reports, consumers are more likely to purchase again on the platform. We have replicated this analysis using a continuous version of relevance and controllability. The results, reported in Appendix Table 7, show the same sign of the estimated effects, but the estimates on $Relevance * Post_{Implement}$ are no longer statistically significant. Taken together, these results suggest 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 7: Consumer Retention

| *Outcome Variable: Share of consumers who buy again in 6 months, times 100* | | |
|---|---|---|
| | (1) | (2) |
| Relevance×Post$_{Implement}$ | 0.159* | 0.157* |
| | (0.086) | (0.086) |
| Control×Post$_{Implement}$ | | 0.015 |
| | | (0.085) |
| Constant | 84.274*** | 84.273*** |
| | (0.012) | (0.017) |
| Observations | 7,147 | 7,147 |
| R-squared | 0.972 | 0.972 |
| Market-seasonality Fixed Effect | ✓ | ✓ |
| Month Fixed Effect | ✓ | ✓ |

*Notes*: One observation is a market-month pair. Outcome variable is the share of consumers who buy again in the following six months in any market, multiplied by 100; $Post_{Implement}$ 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

The policy change increases seller controllability of certification management by removing consumer ratings from the quality certificate requirements. As a result, there is a much reduced variance in the mapping between seller effort and certification status. Given that markets vary in the level of controllability before the policy change, we would expect to see a homogenization in the density of

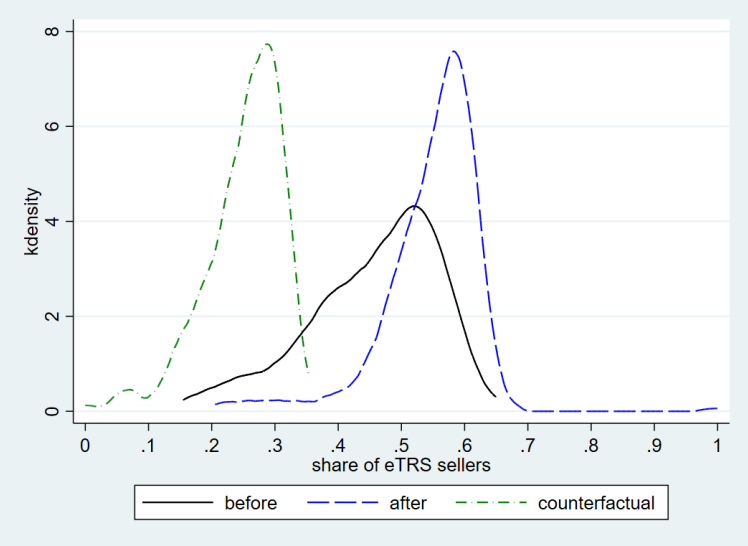
certified sellers across markets after the policy change.

As discussed in Section "Conceptual Framework", depending on the changes in consumer valuation of the new quality certificate, sellers will change their effort, leading to changes in total revenue in the marketplace. Market concentration may also change if the new regime benefits certain seller types. Because of the theoretical ambiguity of the predicted policy effects, we examine them empirically.

### 6.1 Homogenization of Share of Certified Sellers Across Markets

In this section, we show that the distributions of certified sellers in different markets become more similar after the policy change. First, recall that Figure 6 shows consistent evidence of this: Fewer sellers are certified in low-controllability markets in the old regime; the counterfactual shares are more similar across low- and high-controllability markets; in the new regime, the shares of certification are essentially identical across low- and high-controllability markets, except for the largest sellers.

Figure 8: Share of Certified Sellers Across Markets




*Notes*: The solid curve corresponds to the actual share of certified sellers from Month -6 to Month -1. The dash-dotted curve corresponds to the counterfactual share of certified sellers at the policy announcement month (Month -5), normalized by a constant. The dashed curve corresponds to the actual share of certified sellers from Month 0 to Month 5.

To provide more direct evidence on this homogenization, in Figure 8 we plot the distribution of the share of certified sellers across all 336 markets. The solid curve corresponds to the actual share of certified sellers in the months before the regime change, from Month -6 to Month -1. The dashed

curve corresponds to the actual share of certified sellers after the regime change, from Month 0 to Month 5. The dash-dotted curve corresponds to the counterfactual share of certified sellers in the policy announcement month (Month -5), normalized by a constant.[14]

A key takeaway from the figure is that, because the new certification relies on administrative data, the distribution of the counterfactual eTRS shares becomes more concentrated compared to the distribution in the before period and continues to be so in the after period. 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.[15]

### 6.2 Sales and Market Concentration

So far, we have established that changes in certification requirements can affect the selection and behavior of sellers. Based on the theoretical framework discussed in Section "Conceptual Framework", in equilibrium, consumers form new expectation about seller quality, which could affect different seller types differently and, therefore, change the 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 are in a market. Additionally, the regime change could affect seller composition and market concentration, depending on how different seller types benefit from the certification. The question we seek to answer now is: What are the market-level effects of switching from using mainly consumer reports to using administrative data in quality certification?

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

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

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; $\xi_t$ is month fixed effects. We

[14] We have also plotted these densities by months, instead of grouping them, and the results, reported in Appendix Figure 7, are qualitatively the same.

[15] 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.

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.[16] All regressions cluster standard errors by market. The Post dummy corresponds to the policy implementation date, because the change in certification rules became effective only after that date, and this change could in turn affect the market structure. Lastly, note that we use the *lrhc* group (low-relevance-high-controllability) as the baseline;[17] results using the *lrlc* group as the baseline are qualitatively similar and are reported in Appendix Table 7.

Recall from Section "Data" that our sample period contains the 11 months before the policy announcement, the 5 months between the policy announcement and its implementation, and the 6 months after the policy implementation. We start by plotting the times series of the market outcomes for different market types in Appendix Figure 6. We see mostly parallel pre-trends in the outcome variables, particularly in the few months leading up the policy implementation date. The exceptions are HHI (sub-figure d) and the share of sales from small sellers (sub-figure f).

The results are reported in Table 8. There are two high-level findings. First, as columns (1) and (2) show, we do not find statistically significant differences in the changes in sales revenue or volume across the different market types. The lack of significant estimates suggests that the increased controllability under the new regime motivates sellers to increase their effort; otherwise, we might have anticipated a decline in sales revenue and volume due to buyers perceiving the new regime as less relevant.

Second, columns (3) to (6) offer suggestive evidence that markets are becoming more concentrated in low-controllability markets. Specifically, comparing the coefficient estimates for $hrhc \times Post$ and $hrlc \times Post$ reveals that in low-controllability markets, there is a smaller number of sellers achieving sales, increased market concentration (although it's important to note the non-parallel pre-trends), a lower proportion of small sellers among those who have successfully made sales, and a reduced sales share from small sellers (though this is not statistically significant). The qualitative results on market concentration by controllability, based on the coefficient estimates of $lrlc \times Post$—with *lrhc* markets as the baseline—are similar. We have replicated this analysis using a continuous version of relevance and controllability, and report the results in Appendix Table 11.

[16] In the previous analyses on sellers in Section "Results on Sellers" and on buyers in Section "Results on Buyers", we study 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.

[17] The choice of baseline is based on our hypothesized effects of moving from a consumer-rating–based certification to an administrative-data–based certification. Specifically, we expect buyers in high-relevance markets to be more positively affected (or equivalently, less negatively affected) than buyers in low-relevance markets. In addition, we expect sellers in low-controllability markets to be more positively affected than sellers in high-controllability markets.

The key qualitative results stay the same: Specifically, the sales share from small sellers becomes smaller in more relevant markets and in less controllable markets.

Because the regime change reduces the variance in the mapping between seller effort and certification outcome, it should theoretically benefit smaller sellers, who are disproportionately affected by the variance in consumer ratings because of their size. However, we do not have evidence they benefit relative to larger sellers; if anything, we have weak suggestive evidence that smaller sellers are hurt by the regime change. We speculate that the institutional design of quality certificates on eBay could explain this finding: Sellers acquire the eTRS status at the seller level, which potentially benefits all their listings, and the eligibility requirements for the eTRS status include a minimum number of past sales. Both rules increase large sellers' incentive to get the new certification, relative to their smaller counterparts. The observed patterns in market concentration might differ in marketplaces where either certification is given at the listing level or certification requirements place less weight on the volume of past sales.

Table 8: Sales and Market Concentration

| | (1) log(sales revenue USD) | (2) log(sales volume) | (3) log(number sellers w/sales) | (4) log(HHI) | (5) share small sellers among those w/sales | (6) share sales (USD) from small sellers |
|---|---|---|---|---|---|---|
| lrlc×Post | -0.148 | -0.110 | -0.126 | 0.042 | -0.008*** | -0.006 |
| | (0.206) | (0.146) | (0.116) | (0.044) | (0.002) | (0.008) |
| hrhc×Post | -0.015 | -0.022 | -0.014 | 0.006 | -0.005** | 0.002 |
| | (0.038) | (0.025) | (0.014) | (0.048) | (0.002) | (0.009) |
| hrlc×Post | -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; *Post* is defined based on the policy implementation date. We use market-quarter fixed effects (FE) to control for market seasonality. HHI ranges from 0 to 10,000. In parentheses are standard errors clustered at the market level. *** p<0.01, ** p<0.05, * p<0.1.

# 7 Conclusion and Discussion

In this paper, we study the effect of a policy change in quality certification on a leading e-commerce platform whereby administrative data replaced consumer reports in the certification criteria. 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 for obtaining the certification. Additionally, in markets where the new quality certificate is more aligned with consumer experience, buyers place a higher value 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 seem to be 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 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, when product quality can be accurately measured using administrative data on seller behavior, market designers should avoid using consumer reports. For example, in markets for data feeds, which include financial markets, weather data services, and social media analytics, the quality of data feeds can sometimes be objectively assessed by their accuracy and latency. Introducing consumer reports into these markets could add unnecessary variability without substantial benefits. Conversely, in markets where administrative data are a poor predictor of seller quality, often because of greater information asymmetry between the platform and sellers (e.g., the hospitality industry and markets for freelancers), market designers should put more emphasis on consumer reports while relying less on administrative data. Lastly, if market designers prioritize consumer (resp., seller) welfare, they should increase their emphasis on 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 some sellers to improve the focal metrics; on the other hand, better controllability motivates some 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 is a growing theoretical literature on optimal information control (e.g., Vellodi (2018), Saeedi and Shourideh (2019), Shi et al. (2020)), research on the optimal design of noise in certification thresholds is scarce. Our empirical results suggest that this design parameter may deserve more explicit attention in future research.[18]

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 offer weak suggestive evidence that the design of quality certificates could have long-run implications on the marketplace because it could 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 the 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 simultaneously embodies a seller reputation system (i.e., seller ratings based on consumer reports), a seller quality certification program (i.e., 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[19] and (2) eBay's generous buyer warranty program reduces the impact of quality certificates (Hui et al. (2016)). With these caveats in mind, however, we note that multifaceted trust systems are common in e-commerce. For example, 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

[18] On a related note, the choice between a binary certificate vs a continuous (or categorical) measure of quality entails a trade-off between the salience in the eyes of consumers and the threshold effect on the seller side. While we do not study salience directly in this paper, the eTRS premium could be partly driven by that salience. The shirking result of our threshold effect can therefore be seen as a cost that a platform manager should trade off against the benefits of keeping consumer salience with a binary system, and the benefits of increasing controllability and motivating seller efforts with a clearer threshold of the binary certificate.

[19] See Section "Background" for our argument on why this feature is unlikely a big problem empirically in the setting of eBay.

consumer reports separately from these badges. These examples suggest that our results are likely relevant for other digital platforms.

The second limitation is that our findings may not apply to platforms with centralized, standardized services such as warehousing or a uniform dispute resolution system. Instead, the results from our setting are more likely to be generalizable to other firms that rely on agent effort to meet consumers' needs.

Third, 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.

Fourth, we cannot pin down the exact factors driving variation in relevance and controllability across different product markets. Potential explanations encompass the extent of horizontal differentiation, varying levels of information asymmetry between the platform and sellers, or the frequency of purchases in different markets. Nevertheless, we think that the two metrics hold practical value for designing quality certificates, because they are predictive of the effects of changes in certification regimes, easy to measure, and intuitive to managers.

Lastly, our paper cannot quantify welfare changes due to the policy change, which would require a structural model. Nevertheless, 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, because we do not observe the costs of seller effort, nor seller's entry or exit. 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 are 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. On eBay, the existence of the quality certificate does not enter into the search ranking algorithm,[20] but this information may be incorporated in other platforms' search ranking algorithms. The role of certification in consumer search is a topic that warrants future research.

[20]The ranking algorithm is based on a score that depends on the predicted conversion rate of a listing and whether the listing is sponsored. Note that sponsored ads were not introduced on eBay until September 2018.

## 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.

**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.

**Baker, George, Robert Gibbons, and Kevin J Murphy**, "Subjective performance measures in optimal incentive contracts," *The quarterly journal of economics*, 1994, *109* (4), 1125–1156.

**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.

**Bishop, Christopher M and Nasser M Nasrabadi**, *Pattern recognition and machine learning*, Vol. 4, Springer, 2006.

**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.

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

**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.

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

**Einav, Liran, 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.

**Fan, Ying, Jiandong Ju, and Mo Xiao**, "Reputation premium and reputation management: Evidence from the largest e-commerce platform in China," *International Journal of Industrial Organization*, 2016, *46*, 63–76.

**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.

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

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

**Holmström, Bengt**, "Moral hazard and observability," *The Bell journal of economics*, 1979, pp. 74–91.

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

**Hui, Xiang, Maryam Saeedi, 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.

**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.

**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.

**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.

**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.

**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.