When Better Is Worse:

How a Failed Image Quality Experiment Revealed What eBay's Buyers Actually Value

Anjan Goswami, Ph.D.

Abstract

We built an image quality scoring system for eBay product listings, trained on crowdsourced judgments of thousands of product images and validated through click prediction models that showed statistically significant improvements in precision and NDCG when image features were added to the ranking model. The obvious next step—using the quality score to boost listings with better images in search ranking—produced cleaner, more visually appealing search results pages. Revenue dropped. The diagnosis revealed a non-intuitive fact about eBay's core marketplace: in key categories such as used handbags, pre-owned clothing, and vintage goods, "low quality" mobile phone photographs function as **authenticity signals**. A grainy photo of a used Coach bag taken on a kitchen table communicates that the seller is a real person selling a real item they actually own. A studio-quality photograph of the same bag signals professional resale or potential counterfeit. eBay's buyers—particularly in the consumer-to-consumer categories that define the platform's identity—were not optimizing for visual aesthetics. They were optimizing for trust. This case study describes the image quality system, the experiment, the failure, the diagnosis, and the broader lesson: in marketplace search, the definition of "quality" is endogenous to the marketplace's identity, and importing quality standards from adjacent domains (stock photography, retail e-commerce) can actively harm the platform's core value proposition.

1 The Hypothesis: Better Images Should Improve Everything

The logic was straightforward and, on its surface, unassailable. Product search is a visual medium. When a customer scans a search results page, the thumbnail image is the dominant signal driving attention and clicks. A higher-quality image—better lighting, clearer focus, professional composition, clean background—should convey more information about the product, increase buyer confidence, attract more clicks, and ultimately drive more purchases.

This hypothesis was supported by a large body of research in e-commerce and information retrieval. Image quality had been shown to correlate with click-through rates, conversion rates, and perceived product quality across multiple domains. Amazon, the dominant comparison point, had codified this into seller requirements: white backgrounds, professional lighting, multiple angles.

We set out to bring this logic to eBay at scale: build an automated image quality scoring system, identify listings with poor images, and use the quality signal to improve search ranking.

2 Building the Image Quality Scoring System

2.1 Defining Product Image Quality

The first challenge was defining what "quality" means for a product image. Traditional image quality assessment (IQA) focuses on photographic properties: sharpness, noise, exposure, color accuracy. These metrics work well for camera benchmarking and stock photography. They are insufficient for product images.

A product image serves a different purpose than a landscape photograph. It must convey information about a specific item—its condition, its features, its authenticity—in a thumbnail-sized viewport on a search results page. A technically perfect photograph of a product shot from an uninformative angle is worse than a technically imperfect photograph that clearly shows the item's key features.

We defined product image quality along multiple dimensions:

- **Photographic features**: Exposure, sharpness, noise level, colorfulness, contrast—the standard IQA metrics that capture technical image quality
- **Foreground clarity**: Is the product clearly visible and distinguishable from the background? Can the viewer identify what the item is at thumbnail resolution?
- **Background quality**: Is the background clean and non-distracting? Does it compete with the product for visual attention?
- **Composition**: Is the product well-framed? Does it occupy an appropriate proportion of the image? Is it centered or appropriately positioned?
- **Information content**: Does the image convey useful information about the product's condition, features, and characteristics?

The key insight in the definition was that **product image quality is not the same as photographic quality**. A beautifully composed landscape has high photographic quality but zero product information content. A slightly blurry photograph of a circuit board taken under fluorescent lighting has low photographic quality but high information content for an electronics buyer. Our quality model needed to capture both dimensions.

2.2 Crowdsourced Quality Judgments

We conducted a large-scale crowdsourcing experiment to collect human judgments of image quality on thousands of eBay product images. Evaluators classified images into three quality tiers—good, fair, and poor—based on guided perceptual criteria that combined the photographic and product-information dimensions.

The crowdsourced judgments served dual purposes: they provided training labels for the automated quality classifier, and they revealed the distribution of image quality across eBay's catalog. The distribution was sobering: a substantial fraction of listings had images that evaluators rated as "poor"—out of focus, badly lit, cluttered backgrounds, product barely visible. The opportunity for improvement appeared enormous.

2.3 The Automated Quality Classifier

Using the crowdsourced labels, we trained a multi-class classification model to predict image quality from extracted features. The feature set combined low-level photographic features (computed from pixel-level statistics) with higher-level features capturing foreground-background separation, composition, and visual salience.

The classifier achieved strong agreement with human judgments, accurately distinguishing good from poor images across product categories. We published the methodology and results in [1].

3 Image Features Improve Click Prediction

Before using the quality score in ranking, we validated that image quality *mattered* for user behavior. We augmented eBay's standard click prediction model—which used textual features, price, seller reputation, and other standard signals—with two types of image features:

- **Photographic features**: The same features used in the quality classifier—exposure, sharpness, colorfulness, contrast
- **Object features**: Higher-level visual representations learned through convolutional Restricted Boltzmann Machines (RBMs), capturing shape and texture patterns that hand-crafted features could not express

The experiment used eBay search log data spanning one month, restricted to fixed-price, multi-quantity items to control for auction dynamics. Each observation was a click or a skip (an unclicked item appearing above a clicked item in the search results).

The results were unambiguous: **augmenting the click prediction model with image features produced statistically significant improvements in both precision and NDCG.** Image quality was not just correlated with user behavior—it was predictive, incremental to all other features, and the learned object features captured visual signals that photographic features alone could not. We published these results in [2].

The finding seemed to confirm the hypothesis: images matter, better images get more clicks, and we had a validated quality signal ready to deploy in ranking.

4 The Experiment: Boosting High-Quality Images in Search

With a validated quality scorer and evidence that image features predicted clicks, we designed the production intervention: use the image quality score as a ranking signal to boost listings with higher-quality images in search results.

The mechanism was a multiplicative boost in the ranking score: listings with images scored as "good" received a positive adjustment; listings with images scored as "poor" received a negative adjustment. The boost was calibrated to be influential but not dominant—image quality would break ties and shift borderline rankings, not override strong relevance or behavioral signals.

We deployed this as a controlled A/B experiment: treatment users saw search results with the image quality boost applied; control users saw the existing ranking.

4.1 The Results Pages Looked Better

The immediate visual effect was striking. Search results pages in the treatment group were cleaner, more visually cohesive, and more aesthetically appealing. Listings with blurry, dark, or cluttered images were demoted; listings with clear, well-lit, well-composed images were promoted. Side-by-side comparisons of treatment and control pages showed an obvious improvement in visual quality.

By every measure of page aesthetics, the experiment was working.

4.2 Revenue Dropped

The engagement and transaction metrics told a different story. **Revenue in the treatment group declined relative to control.** The experiment that made search results look better was making the marketplace perform worse.

The result was statistically significant and consistent across the experiment duration. This was not noise. Improving image quality in search ranking was actively harming eBay's business.

We did not simply revert the experiment. We investigated.

5 The Diagnosis: Authenticity Trumps Aesthetics

5.1 Category-Level Decomposition

The first analytical step was decomposing the revenue impact by product category. The aggregate revenue decline masked a more nuanced pattern:

- In categories dominated by new, commodity products (electronics accessories, office supplies, new-in-box items), the image quality boost had a **neutral to slightly positive** effect. For these items, a cleaner image was genuinely helpful—it conveyed product information more clearly without adding or removing trust signals.
- In categories dominated by used, pre-owned, and vintage items—**the categories that define eBay's identity and competitive advantage**—the image quality boost had a **significant negative** effect. Revenue dropped substantially in used handbags, pre-owned clothing, vintage watches, collectibles, and similar categories.

The negative effect in eBay's core categories overwhelmed the modest positive effect in commodity categories, producing the aggregate revenue decline.

5.2 What "Low Quality" Images Actually Communicate

We examined the listings that were demoted by the image quality boost in the affected categories. The pattern was immediately recognizable:

A used Coach handbag photographed on a kitchen counter with a mobile phone. Slightly off-center, ambient lighting, a bit of background clutter visible. Our quality classifier scored it "poor." But to an eBay buyer browsing used handbags, this image communicates:

- **This is a real item.** It exists physically in someone's home. It is not a stock photograph copied from the manufacturer's website.
- **The seller is a real person.** Not a professional reseller, not a counterfeiter with a photography studio, but an individual selling something from their closet.
- **The item is authentic.** Counterfeit luxury goods are typically photographed with professional lighting and white backgrounds because the seller is trying to make the fake look legitimate. A casual mobile phone photo is hard to fake in the same way.
- **The item's condition is honestly represented.** The imperfect lighting and composition suggest the seller is not trying to hide flaws or make the item look better than it is.

The "high quality" images that our system promoted in these categories had the opposite effect. A used handbag photographed in a studio with professional lighting and a white background triggered suspicion: *Is this a professional reseller marking up the price? Is this a stock photo—is the actual item different? Is this counterfeit?*

In eBay's core categories, image imperfection was a trust signal. The "low quality" that our automated system penalized was precisely the visual language that eBay's buyers used to assess authenticity.

5.3 Understanding eBay's Core Buyer

The diagnosis forced us to reexamine our assumptions about eBay's user base. eBay is not Amazon. Amazon is a retail platform where customers expect a standardized, professional shopping experience. eBay's competitive advantage is the **consumer-to-consumer marketplace**—real people selling real items to other real people.

eBay's core buyers in C2C categories have developed a sophisticated visual literacy specific to the platform:

- They **prefer mobile phone photos** over professional photography for used items, because mobile photos are harder to fabricate
- They **look for environmental context**—a watch on someone's wrist, a dress hanging in a closet, a bag on a table—as evidence that the seller actually possesses the item

- They **are suspicious of over-polished listings** in categories where items should be unique and individually owned
- They **use image style as a proxy for seller type**: casual photos → individual seller → authentic item → fair price; professional photos → professional reseller → higher markup or potential counterfeit

This visual literacy is rational. In a marketplace with information asymmetry—the seller knows the item's true condition and authenticity, the buyer does not—buyers develop heuristics for assessing seller credibility. On eBay, image style is one such heuristic, and it encodes information about the seller that the image's *content* alone does not convey.

6 Why Click Prediction and Revenue Diverged

A natural question: if image quality improved click prediction (Section 3), why did it harm revenue when deployed in ranking?

The answer is that **clicks and purchases are governed by different decision processes**. Click prediction measures whether a user will click on a search result. Clicking is a low-commitment action driven by visual attention and curiosity. A high-quality image does attract more clicks—the click prediction results were correct.

But **purchasing requires trust**, especially for high-value used items. A buyer might click on a professionally photographed used Chanel bag out of curiosity, then decide not to purchase because the listing *feels* like a professional reseller or potential counterfeit. Conversely, a buyer might click on a casually photographed listing and purchase with confidence because the image communicates authenticity.

The image quality boost increased clicks on high-quality images (improving CTR) while decreasing purchases from those clicks (reducing conversion), and simultaneously decreased clicks on "low quality" images that had higher conversion rates. The net effect on revenue was negative because the conversion effect dominated the click effect in eBay's high-value C2C categories.

This is a specific instance of a general problem: **optimizing for an intermediate metric (clicks) can harm the downstream metric (revenue) when the intermediate metric is not aligned with the factors that drive the downstream decision.**

7 Implications and Generalizable Principles

7.1 Quality Is Endogenous to the Marketplace

The most important lesson is that **there is no universal definition of quality**. What constitutes a "good" product image depends on the marketplace, the category, the buyer population, and the competitive dynamics. Importing quality standards from Amazon (white backgrounds, professional lighting) into eBay's C2C marketplace was not an improvement—it was a misunderstanding of what eBay's buyers value.

This principle extends beyond images. Any quality signal—review scores, seller ratings, listing completeness—must be defined relative to the marketplace's identity and buyer expectations. A five-star seller rating means something different on eBay ("this individual was honest about the item's condition") than on Amazon ("this retailer shipped quickly and the item matched the product page").

7.2 Information Asymmetry Creates Non-Obvious Quality Signals

In markets with information asymmetry, signals that appear to indicate low quality may actually indicate high credibility. This is a well-known phenomenon in economics (the concept of "costly signaling"), but it is rarely incorporated into ML quality scoring systems.

A casual mobile phone photo is a costly signal in the sense that a counterfeiter or dishonest seller is unlikely to produce it—they would invest in professional photography to make their listing more attractive. The "cost" of the casual photo is the aesthetic penalty; the "signal" is authenticity. Buyers who have learned to read this signal use it rationally.

ML systems that score quality based on photographic properties will systematically penalize these costly signals because they appear as defects rather than information. **Any quality scoring system deployed in a marketplace with information asymmetry must be validated against transaction outcomes, not just perceptual judgments.**

7.3 Failed Experiments Are the Most Valuable Experiments

The image quality experiment "failed" in the narrow sense that it did not produce the intended revenue lift. It succeeded in a deeper sense: it revealed a fundamental property of eBay's buyer behavior that no amount of log analysis, user research, or survey data would have uncovered with the same clarity.

Before the experiment, the team's mental model of eBay's buyers was implicitly calibrated to a retail paradigm: buyers want clean, professional shopping experiences, and higher production values lead to better outcomes. After the experiment, the mental model was recalibrated: eBay's buyers in core categories are **trust-seeking, not aesthetics-seeking**, and they have developed a sophisticated visual literacy that uses image style as a proxy for seller credibility.

This recalibration influenced subsequent ranking decisions, seller guidance, and category-specific strategies. The "failed" experiment produced more strategic value than many successful ones.

7.4 Aggregate Metrics Mask Heterogeneous Effects

The aggregate revenue decline masked a category-level pattern that was both more nuanced and more actionable. In commodity categories, image quality improvement was benign or positive. In C2C categories, it was harmful. A more granular experiment design—stratified by category type—would have revealed this heterogeneity before the aggregate result created organizational concern.

This is a general lesson for marketplace experimentation: **always decompose by category, buyer segment, and seller type before drawing conclusions from aggregate metrics.** The average treatment effect in a diverse marketplace is often meaningless because it averages over subpopulations with opposite responses.

8 Published Work

The image quality assessment methodology and the click prediction results were published in two peer-reviewed venues:

1. **Goswami, A., Chung, S. H., Chittar, N., and Islam, A.** "Assessing product image quality for online shopping." *Proc. SPIE 8293, Image Quality and System Performance IX*, 82930L (2012). DOI: 10.1117/12.906982

2. **Chung, S. H., Goswami, A., Lee, H., and Hu, J.** "The impact of images on user clicks in product search." *Proceedings of the Twelfth International Workshop on Multimedia Data Mining (MDMKDD '12)*, ACM (2012). DOI: 10.1145/2343862.2343866

References

[1] A. Goswami, S. H. Chung, N. Chittar, and A. Islam. Assessing product image quality for online shopping. In *Proc. SPIE 8293, Image Quality and System Performance IX*, 82930L, 2012.

[2] S. H. Chung, A. Goswami, H. Lee, and J. Hu. The impact of images on user clicks in product search. In *Proceedings of the Twelfth International Workshop on Multimedia Data Mining (MDMKDD '12)*, pages 25–33. ACM, 2012.