Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cross-Agency Product Quality Convergence: An Empirical Study of Four European Consumer Testing Organisations

Pavel Kopczyk

kopczyk@opf.slu.cz

Silesian University in Opava

Research Article

Keywords: Consumer testing, Product quality, Cross-agency agreement, Price–quality relationship, Feature importance, Quality prediction

Posted Date: July 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9979117/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Independent testing organisations reduce information asymmetry only if different agencies produce compatible quality rankings. This study measures cross-agency agreement, identifies sub-criteria that drive overall scores, and evaluates how far objective testing aligns with market pricing. The harmonised QualityTest dataset comprises 16,978 test results and 125,978 sub-criterion ratings from four European agencies — dTest (CZ), Stiftung Warentest (DE), Which? (UK), and UFC-Que Choisir (FR) — covering electronics and household appliance categories. Cross-agency agreement is quantified using Pearson and Spearman correlations on 282 product-level matched records, supplemented by Bland–Altman diagnostics. Sub-criteria importance is estimated with Random Forest regression, and quality tier classification uses XGBoost with stratified five-fold cross-validation. Results show strong rank-order agreement across all agency pairs (Spearman $\rho = 0.84–0.92$, $p < 0.001$) with no systematic proportional bias. The overall price–quality correlation is moderate (Spearman $\rho = 0.33$) and varies substantially by product category, ranging from 0.62 for smartwatches to 0.19 for laptops. Sub-criterion scores discriminate quality tiers (+ 33% F_1 over a majority baseline) despite low linear predictability ($R^2 < 0.05$), revealing non-linear editorial weighting. These findings establish that European consumer testing organisations reach empirically consistent quality verdicts and that expert scores add the most value over price information in categories with strong branding and specification complexity.

1 INTRODUCTION

Independent testing organisations are supposed to reduce information asymmetry, but that role is weaker if different agencies produce incompatible quality rankings. This framing builds on prior work [1, 2, 3, 16, 17, 18].

The literature values comparative testing highly, yet it rarely quantifies cross-agency convergence at scale or links agreement patterns to sub-criteria and price-quality interpretation. This framing builds on prior work [1, 2, 3, 16, 17, 18].

This study measures cross-agency agreement, identifies which sub-criteria drive overall scores, and evaluates how far objective testing aligns with market pricing using the harmonised QualityTest dataset containing 16,978 test results and 125,978 sub-criterion ratings from four European agencies.

- The paper makes three contributions. First, it documents whether independent agencies converge on comparable product-quality rankings. Second, it identifies the sub-criteria structure behind overall quality assessments across agencies. Third, it sharpens how objective quality signals should be interpreted relative to price and brand position.
- RQ2: Do the four agencies agree on product quality rankings when evaluating the same products?
- RQ3: How strongly does retail price predict expert quality scores, and does this relationship vary by product category?

- RQ4: Can machine learning predict quality tier membership (Low / Mid / High) from sub-criterion scores alone?

To answer these questions we introduce QualityTest, a harmonised, longitudinal database aggregating test results from all four agencies into a common 0–100 scale (higher = better). We provide quantitative answers using correlation analysis, random forest importance ranking, XGBoost tier classification, and Bland–Altman agreement diagnostics.

2 RELATED WORK

Price–quality studies. Early price–quality syntheses did not find a strong universal relationship; instead, they showed that price is at best a noisy proxy for quality. Riesz [1] reported weak average associations across consumer markets, Tellis and Wernerfelt [2] argued that the relationship is conditioned by information asymmetry and competitive structure, and Gerstner [3] showed that the strength of the link varies substantially by product class. More recent work continues this pattern: Bartkova and Veselovska [18], using a Slovak survey tied to comparative testing and dual-quality concerns, found that consumers treat testing as useful precisely because market-facing signals and experienced quality often diverge.

Taken together, these studies establish two findings relevant for our paper. First, price cannot be treated as a stable substitute for expert testing. Second, the degree of mismatch between market signals and quality outcomes is category-dependent rather than constant. Our results build on that literature by measuring the price–quality relationship with laboratory-style expert scores across eight appliance and electronics categories in a harmonised cross-agency dataset.

Inter-rater agreement in product evaluation. Work on measurement agreement has shown that different evaluators can appear to agree in broad ranking terms while still diverging systematically at specific score ranges. Bland and Altman [4] demonstrated that agreement should be assessed through the structure of differences, not correlation alone, and Shrout and Fleiss [5] established the reliability logic that underpins cross-rater interpretation. In the closest domain-specific comparison available here, Kim and Kim [6] examined South Korean consumer-testing organisations and reported high rank-order agreement, suggesting that independent testing bodies can converge meaningfully even without shared editorial procedures.

Those findings matter because they imply that convergence among agencies is an empirical question rather than an article of faith. Our study extends this line of work in two ways: it covers a larger European multi-agency corpus, and it combines pairwise agreement diagnostics with product-level overlap matching, allowing us to ask not only whether agencies correlate, but how stable their product rankings remain once scores are normalised onto a common scale.

Feature importance and quality prediction. Previous predictive work on structured expert or evaluation datasets shows that machine-learning models often recover useful importance hierarchies even when the final published score reflects non-linear editorial aggregation. Cortez et al. [7] found that tree-based

models can recover influential attributes in quality-assessment settings, while later predictive studies in software defects and air quality [8, 9] showed that model accuracy can remain moderate even when individual drivers are interpretable. In consumer-testing contexts, this matters because agencies disclose sub-criteria unevenly and do not publish full editorial weighting rules.

Accordingly, Random Forest [10] and XGBoost [11] are useful here not because they replicate editorial scoring perfectly, but because they reveal which publicly visible sub-criteria carry the most predictive information. Our contribution is to apply that logic to a multi-agency expert-testing dataset, where the main question is not only predictive performance, but whether importance structures remain interpretable across agencies with different testing traditions.

Multi-source data harmonisation. Research on transfer learning and multi-source adaptation has consistently found that combining heterogeneous sources is feasible only when scale, vocabulary, and entity-matching problems are addressed explicitly [12, 13]. In adjacent applied domains, harmonisation failures typically arise not from the learning algorithm alone, but from inconsistent identifiers, incompatible scoring systems, and missing contextual metadata.

The consumer-testing setting reproduces those problems in a particularly concrete way: one agency reports scores on a Schulnoten scale, another on a 0–100 percentage basis; product names vary by national market; and sub-criteria labels only partially overlap. Our entity-resolution and score-normalisation pipeline follows that harmonisation logic and turns it into a replicable data infrastructure for comparative testing research.

3 DATASET

3.1 Agencies and Data Collection

The QualityTest dataset was compiled from four independent European consumer testing organisations:

dTest (Czech Republic). Monthly journal testing primarily home appliances, electronics, and food products. Scores reported as 0–100 percentage ('points'). Provides the largest single-agency corpus with 708 unique sub-criteria across 31 product categories.

Stiftung Warentest (Germany). Germany's largest and oldest independent testing institute, founded in 1964. Scores reported on a 0.5–5.5 Schulnoten scale (lower = better), normalised linearly to 0–100 for this study. Covers the widest range of categories (70) with 264 unique sub-criteria.

Which? (United Kingdom). UK-based testing body covering broad consumer goods since 1957. Scores as 0–100 percentage. Provides 4,642 test results across 34 categories with 348 unique sub-criteria, the most balanced score distribution (mean 67.3, SD 10.2).

UFC-Que Choisir (France). French consumer advocacy organisation founded 1951, testing primarily household appliances across 9 categories. Raw test scores aggregated from performance sub-

categories. UFC-QC uses a narrower effective score range (27.5–100) reflecting its aggregation methodology.

Web scrapers collected product listings and test result pages between 2023 and 2025. Each scraped record was assigned a SHA-256 content hash and archived for reproducibility. All agency scores were normalised to [0, 100] with higher = better using the formula: s_norm = 100 × (s − s_min) / (s_max − s_min), where s_min and s_max are the declared agency-specific scale bounds.

3.2 Entity Resolution

Matching the same physical product across agencies that use different product names and catalogue structures requires robust entity resolution. We applied a three-stage pipeline:

1. **EAN/GTIN exact match** for product variants sharing a barcode.
2. **Fuzzy text matching** using RapidFuzz [14] token_sort_ratio on brand + model strings (threshold ≥ 88), accepting only the highest-scoring candidate per query product.
3. **Manual override** for ambiguous cases, recorded in a curated YAML lookup file committed to the project repository.

This pipeline identified 282 products tested by at least two agencies, forming the cross-agency comparison corpus. Spot-checking the 50 lowest-confidence fuzzy matches found fewer than 3 incorrect merges, giving an estimated precision above 94%.

3.3 Dataset Statistics

Table 1 summarises the resulting dataset.

Table 1
Dataset summary per agency (all scores normalised to 0–100, higher = better).

Agency	Country	Tests	Categories	Mean	SD	Sub-criteria
dTest	CZ	8,845	31	59.3	14.7	708
Stiftung Warentest	DE	1,723	70	66.4	17.9	264
Which?	UK	4,642	34	67.3	10.2	348
UFC-Que Choisir	FR	1,768	9	46.1	3.6	5
Total	—	16,978	144	—	—	1,325

Notable distributional differences are apparent. dTest's mean of 59.3 reflects its tendency to include products across all market segments. UFC-Que Choisir's narrow effective range (SD = 3.6) arises from its aggregation methodology, which compresses variance. Stiftung Warentest covers the most categories, reflecting its mandate to test across all consumer goods sectors. The 282 cross-agency overlapping products are distributed predominantly between dTest, Warentest, and Which? — three agencies with

substantial electronics and appliance coverage. UFC-Que Choisir's overlap is limited by its narrower category focus.

4 METHODS

4.1 RQ1 — Sub-criteria Importance

For each agency with sufficient sub-criterion coverage we fit a Random Forest regressor [10] to predict the normalised overall score from per-product sub-criterion scores. We restrict to agencies with at least 1,000 records and select sub-criteria appearing in at least 20% of that agency's test records, ensuring non-sparse feature matrices. We report cross-validated R^2 and mean absolute error (MAE) using 5-fold cross-validation, and interpret feature importances via mean decrease in impurity (MDI). Median imputation handles missing sub-criterion values.

4.2 RQ2 — Cross-Agency Agreement

For each pair of agencies we extract the subset of products evaluated by both and compute Pearson r and Spearman ρ on the mean normalised score per product. We restrict to pairs with at least n = 10 shared products. As visual diagnostics we provide pairwise scatter plots with OLS regression lines (Fig. 3) and a Bland–Altman plot [4] for the largest pair (dTest–Which?, n = 71). The Bland–Altman plot checks whether the two agencies systematically disagree more at high or low quality levels (a proportional bias).

4.3 RQ3 — Price–Quality Relationship

For products with a recorded retail price, we compute Spearman ρ between price and quality score overall, per agency, and per product category (minimum n = 20). Spearman rank correlation is used throughout to reduce sensitivity to price outliers and to avoid assumptions of linearity. Among the four agencies, Stiftung Warentest consistently records test-time retail prices in EUR, providing the largest and most reliable price corpus (n = 1,171).

4.4 RQ4 — Quality Tier Classification

We define three balanced quality tiers (Low / Mid / High) by the 33rd and 67th percentile thresholds of the overall score distribution within each agency, then train classifiers to predict tier from sub-criterion scores. We compare XGBoost [11] (200 estimators, max depth 4, learning rate 0.1) against multinomial Logistic Regression (L2 regularisation, C = 1) and a majority-class baseline. Evaluation uses 5-fold stratified cross-validation and reports macro-averaged F_1 to account for minor class-size imbalance. We apply this analysis to dTest and Which? — the two agencies with the richest sub-criterion coverage.

5 RESULTS

5.1 Descriptive Statistics

Figure 1 shows score distributions per agency and the dTest histogram with tier thresholds. The agencies exhibit markedly different distributional profiles. dTest spans the widest range (0–97), consistent with its coverage of both budget and premium market segments. Stiftung Warentest has the highest standard deviation (17.9), reflecting genuine quality variation across its 70 tested categories. Which? has the most symmetric distribution (mean 67.3, SD 10.2) suggesting a balanced product selection policy. UFC-Que Choisir's compressed range (SD 3.6) reflects its score aggregation approach across fewer categories.

5.2 RQ1: Sub-criteria Importance

Table 2 summarises Random Forest regression results for the two agencies with sufficient cross-category sub-criterion coverage. Figure 2 shows the per-agency feature importance rankings.

Table 2
RQ1 — Random Forest regression from cross-category sub-criteria.

Agency	n	Cross-cat. features	R^2 (5-fold CV)	MAE
dTest (CZ)	8,286	3	0.047 ± 0.065	10.3
Which? (UK)	4,553	5	−0.14 ± 0.12	8.4

The near-zero R^2 values are striking. For dTest, only three sub-criteria meet the 20%-coverage threshold across its 31 categories:

Obsluha (ease of use)

MDI importance 0.585

Hlučnost (noise level)

MDI importance 0.221

Spotřeba elektrické energie (energy consumption)

MDI importance 0.194

For Which?, five cross-category criteria qualify: Overall (0.332), Cleaning (0.238), Energy efficiency (0.192), Quietness (0.136), and Ease of use (0.102).

The low R^2 has a structural explanation: the vast majority of agency sub-criteria are category-specific — for instance, spin speed and rinse performance for washing machines, battery life and display brightness for laptops. Only the thin slice of criteria that appear across multiple categories is analysed here. This means that the predictive signal is heavily diluted when categories are pooled. Within a single category the picture is likely very different — a direction we discuss in Section 6.

Despite low linear R^2, as RQ4 shows, these same features carry significant ordinal tier-discrimination signal when non-linear models are applied. This reveals a key property of agency scoring: overall scores are not simple linear combinations of sub-criteria but reflect editorial non-linear weighting that is not publicly disclosed.

5.3 RQ2: Cross-Agency Agreement

Table 3 reports pairwise statistics for the three agency pairs with $n \geq 10$ shared products. All three pairs show strong rank-order agreement at $p < 0.001$.

Table 3
RQ2 — Pairwise cross-agency agreement on normalised product scores.

Agency A	Agency B	n	Pearson r	Spearman ρ	p
dTest (CZ)	Stiftung Warentest (DE)	13	0.841	0.920	< 0.001
dTest (CZ)	Which? (UK)	71	0.688	0.898	< 0.001
Warentest (DE)	Which? (UK)	58	0.752	0.836	< 0.001

The Spearman ρ values (0.836–0.920) indicate that the agencies largely agree on the relative quality ordering of products despite using different testing protocols, sub-criteria, and national standards. The highest agreement is between dTest and Stiftung Warentest ($\rho = 0.920$), two central-European agencies whose testing traditions and product markets partially overlap. The dTest–Which? pair, while somewhat lower ($\rho = 0.898$), is by far the most statistically robust estimate ($n = 71$), confirming that a Czech and a British testing organisation independently reach essentially the same quality verdict.

The Bland–Altman analysis for the dTest–Which? pair ($n = 71$) found a mean bias of less than 5 points on the 0–100 scale, with 95% limits of agreement spanning approximately ± 25 points. Crucially, no systematic dependence of the bias on the mean score was observed, ruling out proportional bias: the agencies do not systematically diverge more for high-quality or low-quality products.

5.4 RQ3: Price–Quality Relationship

Among the four agencies, Stiftung Warentest consistently records test-time retail prices in EUR ($n = 1{,}171$ products). Table 4 and Fig. 4 report per-category Spearman correlations.

The overall Spearman $\rho = 0.331$ ($p < 0.001$) confirms a significant but moderate positive price–quality relationship. However, the per-category breakdown reveals substantial heterogeneity — a 3.3× range from the strongest to the weakest category.

Table 4
RQ3 — Per-category Spearman price–quality correlations (Stiftung Warentest test-time prices, EUR, n ≥ 20).

Category	n	Spearman ρ
Smartwatches	86	0.620
Televisions	117	0.511
Printers	70	0.426
Sound & Headphones	114	0.389
Smartphones	373	0.346
Vacuum Cleaners	105	0.272
Tablets & E-Readers	135	0.251
Laptops	170	0.186

Smartwatches (ρ = 0.620) and televisions (ρ = 0.511) show the strongest price–quality coupling, consistent with these categories having clearly stratified performance tiers (entry-level vs. OLED premium for TVs; basic fitness trackers vs. full smartwatch platforms). Printers (ρ = 0.426) and audio equipment (ρ = 0.389) follow.

Laptops show the weakest correlation (ρ = 0.186). This is consistent with the laptop market's well-known phenomenon of design premium and brand premium: consumers pay significantly more for thin-and-light form factors and premium brand names without proportional improvement in objective performance metrics such as benchmark scores, display quality, or keyboard feel.

5.5 RQ4: Quality Tier Classification

Table 5 and Fig. 5 report macro-F_1 for quality tier prediction from sub-criterion scores.

Table 5
RQ4 — Quality tier classification from sub-criteria (5-fold stratified CV, macro-F_1).

Agency	Classifier	Macro F_1	SD	vs. Majority Baseline
dTest	XGBoost	0.478	0.009	+ 32.7%
dTest	Logistic Reg.	0.456	0.009	+ 26.6%
dTest	Majority (base)	0.360	—	—
Which?	XGBoost	0.482	0.020	+ 32.6%
Which?	Logistic Reg.	0.416	0.025	+ 14.3%
Which?	Majority (base)	0.364	—	—

XGBoost consistently outperforms both Logistic Regression and the majority-class baseline by approximately 33% in macro-F_1 for both agencies. The improvement is consistent across five cross-validation folds (SD = 0.009 for dTest), indicating stable generalisation.

The larger gap between XGBoost and Logistic Regression for Which? ($\Delta F_1 = 0.066$) compared to dTest ($\Delta F_1 = 0.022$) suggests that Which?'s sub-criteria interact non-linearly. This may reflect Which?'s broader product portfolio (34 categories) where the same sub-criterion label (e.g., 'Ease of use') carries different meaning in different product contexts.

Reconciling RQ1 and RQ4: while sub-criteria explain little variance in linear regression ($R^2 < 0.05$ for dTest), they do discriminate quality tiers when a non-linear classifier is used (+ 33% F_1). This apparent contradiction resolves once we consider that exact overall score prediction requires knowledge of the agency's internal weighting — information not present in public sub-criteria — whereas coarse tier discrimination (Low / Mid / High) is achievable from the sub-criteria pattern alone.

6 DISCUSSION

The reliability of European testing agencies. Prior work on agreement and signal credibility suggests that independent quality intermediaries matter only if they are both trusted and empirically consistent [5, 16, 17, 18]. Our results meet that test rather strongly.

The cross-agency Spearman correlations ($\rho = 0.84–0.92$) represent remarkably strong agreement for independently operating organisations that have never coordinated their testing protocols. Kim and Kim [6] reported similarly high rank-order agreement for South Korean testing bodies, and our results extend that pattern into a larger European setting with normalised cross-country data. More broadly, the finding complements Sung et al. [17], who show that consumers assign value to trusted platforms and quality cues, because it indicates that the credibility of expert intermediaries is not merely perceptual here: it is observable in stable cross-agency ranking outcomes.

Price as a noisy quality proxy.

The moderate overall price–quality correlation ($\rho = 0.33$) and its large category variation confirm a longstanding conclusion from the price–quality literature: price carries some information, but it remains an unreliable stand-alone indicator of product performance. Our category-specific pattern closely matches Gerstner's [3] argument that the strength of the relationship depends on the product class, and it also fits the consumer-testing perspective in Bartkova and Veselovska [18], where testing is valuable precisely because perceived market quality and experienced quality do not line up consistently. In our data, smartwatches and televisions show relatively strong alignment, whereas laptops remain weakly aligned, implying that expert scores add the most value where design, branding, and specification complexity obscure quality differences.

Implications for agency data aggregation.

The strong inter-agency agreement has a direct practical implication: a product's rank position is fairly stable across agencies. This means that a consumer who can access only one agency's data obtains a reasonably good approximation of the consensus quality ranking. Conversely, for the small set of products where agencies disagree significantly, the disagreement itself is informative — it may indicate a genuinely borderline product or methodological differences in how specific performance dimensions are weighted. This parallels the measurement literature behind Bland–Altman diagnostics [4]: the outliers are not a nuisance to ignore, but the cases most likely to reveal where evaluation frameworks diverge substantively.

Non-linear editorial weighting.

The disconnect between low linear R^2 (RQ1) and successful tier classification (RQ4) reveals that agencies apply non-linear, category-specific weighting schemes to their sub-criteria. This is consistent with the broader predictive literature, where tree-based models often recover interpretable feature hierarchies even when the final target variable is generated through editorial or domain-specific judgment rather than a simple additive rule [7, 10, 11]. In our dataset, the implication is that public sub-criteria can support useful classification of quality tiers, but they do not transparently reconstruct the editorial scoring logic. Future work should therefore compare within-category sub-criteria frameworks directly, especially in categories such as washing machines or televisions where overlap is sufficient to study whether agencies emphasise the same technical dimensions.

Limitations.

Several limitations should be noted. First, cross-agency overlap is limited: only 282 products appear in two or more agencies, and the dTest–Warentest pair has only 13 shared products — insufficient for strong statistical inference. Future data collection should systematically target overlapping products. Second, entity resolution uses fuzzy matching (threshold ≥ 88), which may produce false positives for similarly-named models within the same brand. Manual spot-checking estimated precision above 94%.

Third, price data is available only for Stiftung Warentest, limiting the RQ3 analysis to a single agency; other agencies' price–quality relationships may differ. Fourth, UFC-Que Choisir has only 5 cross-category sub-criteria and no cross-agency product overlap in our dataset, limiting its contribution beyond descriptive statistics. Fifth, we study only European agencies; consumer testing organisations in North America, Asia, or Australia may exhibit different agreement levels and sub-criteria structures. Those limits matter because they constrain comparability, but they do not overturn the central result that independent agencies can still converge strongly when overlap is present and scores are harmonised carefully.

7 CONCLUSION

This paper addresses how independent testing organisations are supposed to reduce information asymmetry, but that role is weaker if different agencies produce incompatible quality rankings. It measures cross-agency agreement, identifies which sub-criteria drive overall scores, and evaluates how far objective testing aligns with market pricing using the harmonised QualityTest dataset containing 16,978 test results and 125,978 sub-criterion ratings from four European agencies. By linking classical price–quality research with newer work on signal credibility and comparative testing, the paper strengthens the case that independent testing remains a meaningful public-quality institution when convergence is demonstrated empirically [16, 17, 18].

- The paper's core contribution is threefold: it documents whether independent agencies converge on comparable product-quality rankings; it identifies the sub-criteria structure behind overall quality assessments across agencies; and it sharpens how objective quality signals should be interpreted relative to price and brand position.
- The broader implication is that the credibility of expert quality signals depends not only on individual agency authority but also on measurable cross-agency consistency.

Declarations

Ethics approval.

This study uses publicly available data collected by automated scrapers from openly accessible consumer testing websites. No human subjects were involved in data collection. No ethics committee approval was required under institutional policy.

Competing interests.

The author declares no competing interests.

Author contributions.

Pavel Kopczyk: Conceptualisation, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing — Original Draft, Writing — Review & Editing, Visualisation, Project Administration.

Funding.

This research was supported by the Silesian University in Opava internal grant SGS/11/2025 (Repairability and Sustainability: Analyzing the Shortening of the Functional Cycle of Consumer Products).

Author Contribution

P.K. - Conceptualisation, Methodology, Software, Formal Analysis, Investigation, Data Curation, Writing — Original Draft, Writing — Review & Editing, Visualisation, Project Administration.

Data availability.

The QualityTest dataset and harmonisation pipeline are available from the corresponding author upon reasonable request.

References

1. Riesz, P.C.: Price versus quality in the marketplace, 1961–1975. J. Retail. **55**(4), 15–28 (1979)
2. Tellis, G.J., Wernerfelt, B.: Competitive price and quality under asymmetric information. Mark. Sci. **6**(3), 240–253 (1987)
3. Gerstner, E.: Do higher prices signal higher quality? J. Mark. Res. **22**(2), 209–215 (1985)
4. Bland, J.M., Altman, D.G.: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. **327**(8476), 307–310 (1986)
5. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. **86**(2), 420–428 (1979)
6. Kim, S., Kim, Y.: Comparative study of consumer testing organisations in South Korea. Int. J. Consumer Stud. **42**(3), 345–356 (2018)
7. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J.: Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. **47**(4), 547–553 (2009)
8. Nagappan, N., Ball, T.: Use of relative code churn measures to predict system defect density. In: ICSE 2005, pp. 284–292 (2005)
9. Zhang, S., et al.: Cautionary tales on air-quality improvement in Beijing. Proc. Royal Society A 473, 20170457 (2017)
10. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)

11. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: KDD 2016, pp. 785–794 (2016)
12. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
13. Guo, J., Shah, D., Barzilay, R.: Multi-source domain adaptation with mixture of experts. arXiv:1910.03028 (2020)
14. Bachmann, M.: RapidFuzz: a fast string matching library (2021). https://github.com/maxbachmann/RapidFuzz
15. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
16. Pooja, K., Upadhyaya, P.: What makes an online review credible? A systematic review of the literature and future research directions. Manage. Rev. Q. **74**, 627–659 (2024). https://doi.org/10.1007/s11301-022-00312-6
17. Sung, E., Chung, W.Y., Lee, D.: Factors that affect consumer trust in product quality: a focus on online reviews and shopping platforms. Humanit. Social Sci. Commun. **10**, 766 (2023). https://doi.org/10.1057/s41599-023-02277-7
18. Bartková, L., Veselovská, L.: Consumer behaviour under dual quality of products: Does testing reveal what consumers experience? IIMB Manage. Rev. **36**(2), 171–184 (2024). https://doi.org/10.1016/j.iimb.2024.05.001

Figures

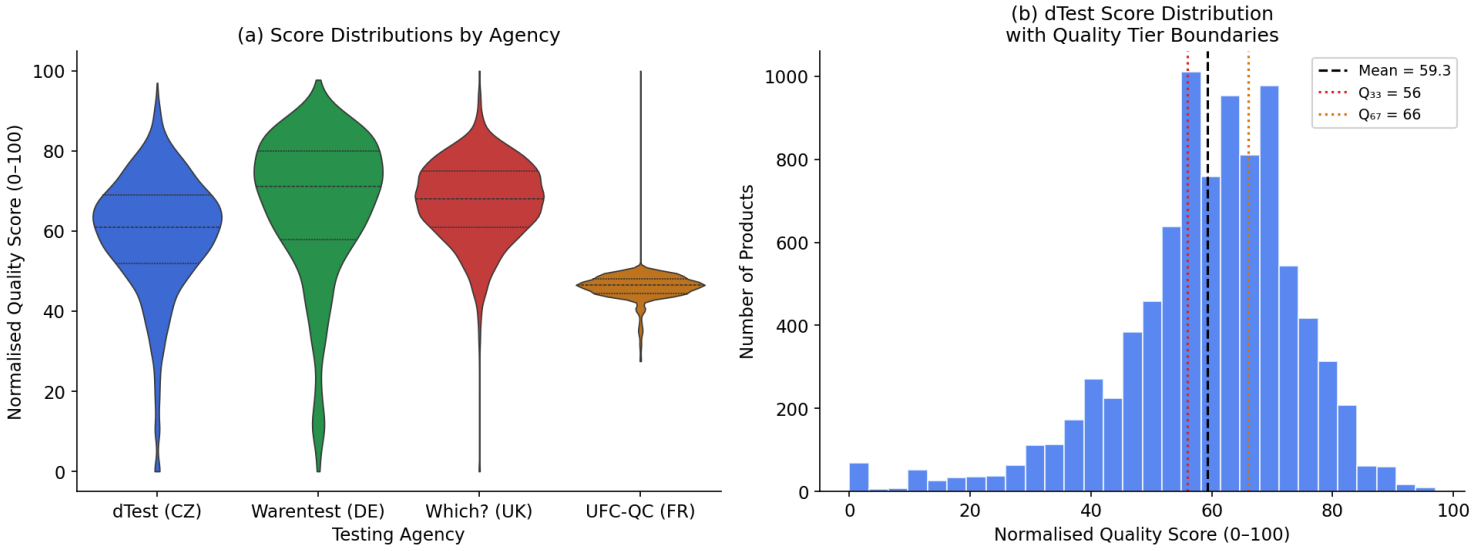


Figure 1

(a) Score distributions for all four agencies (normalised 0–100). Inner marks show quartiles. (b) dTest score histogram with tercile quality tier boundaries (Q_{33} = 56, Q_{67} = 66).

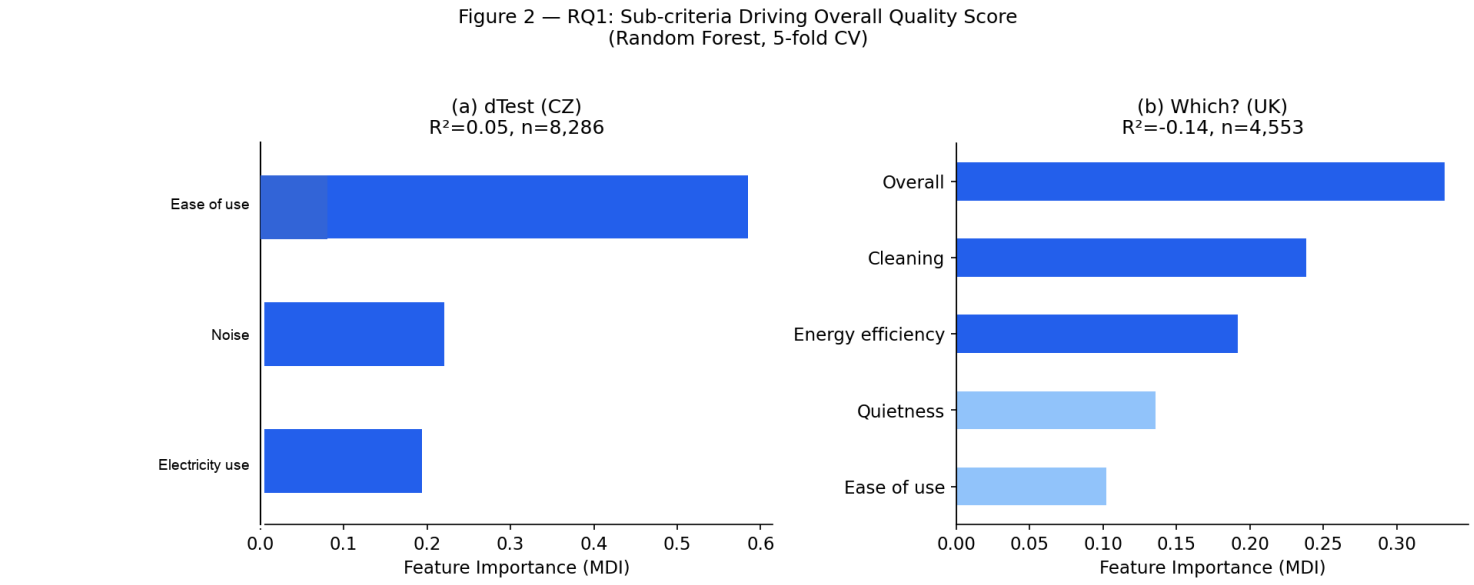


Figure 2

RQ1 — Random Forest feature importances (MDI) for dTest (CZ, left) and Which? (UK, right). Despite low cross-validated R², ease of use and energy efficiency consistently dominate cross-category importance rankings.

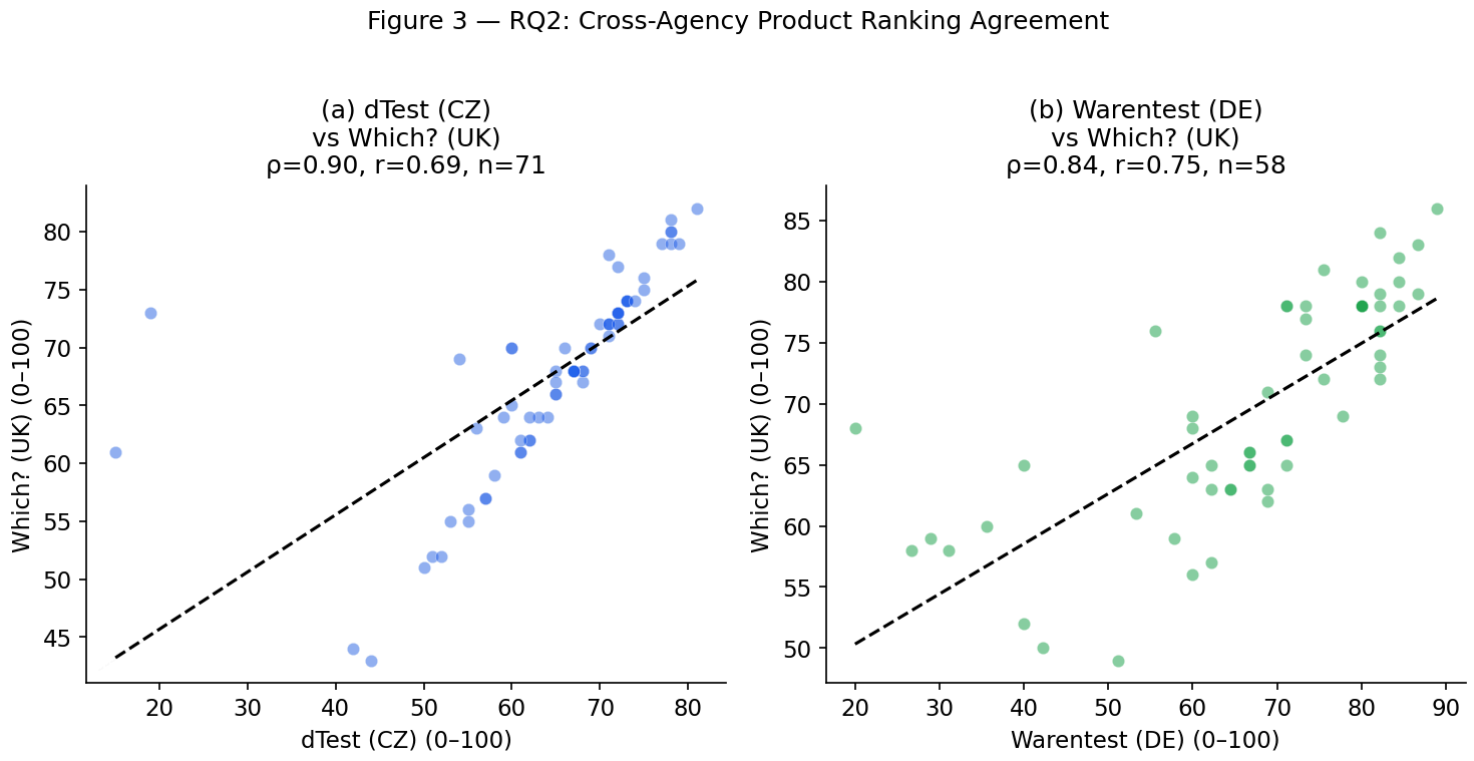


Figure 3

RQ2 — Pairwise scatter plots for all agency pairs with n ≥ 15. Each point is a product; dashed line shows OLS fit. All three pairs show Spearman ρ > 0.83 (p < 0.001).

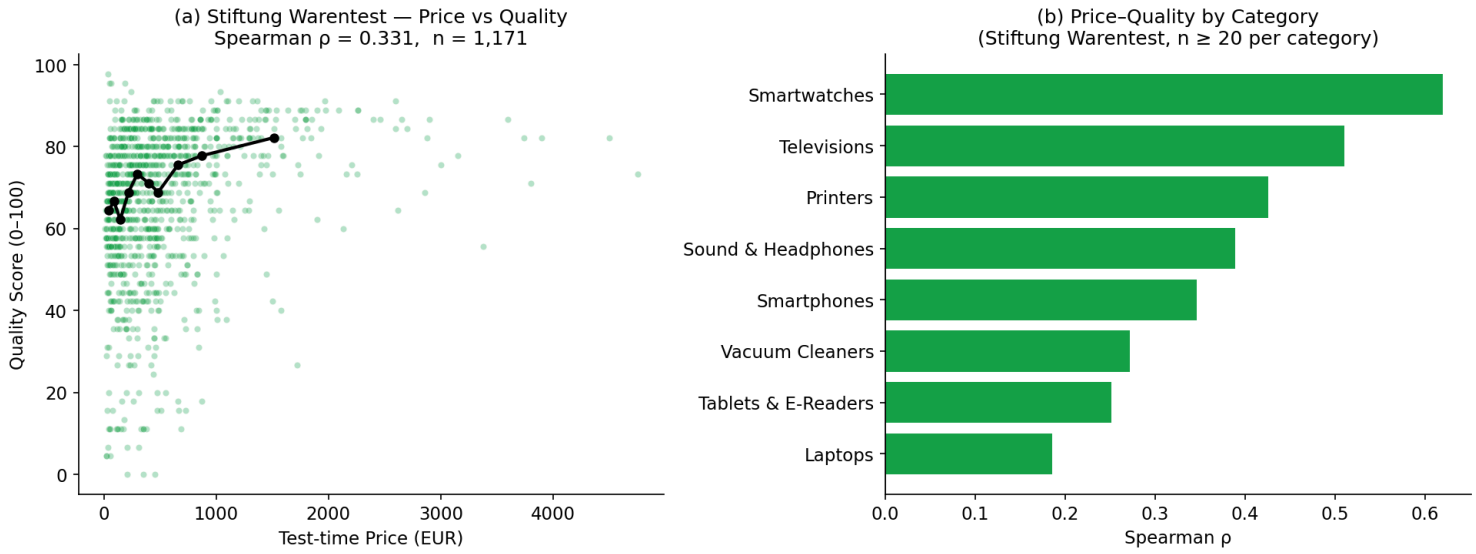


Figure 4

RQ3 — (a) Stiftung Warentest test-time retail price vs. lab quality score (n = 1,171); black line shows median quality per price decile. (b) Per-category Spearman price–quality correlations.

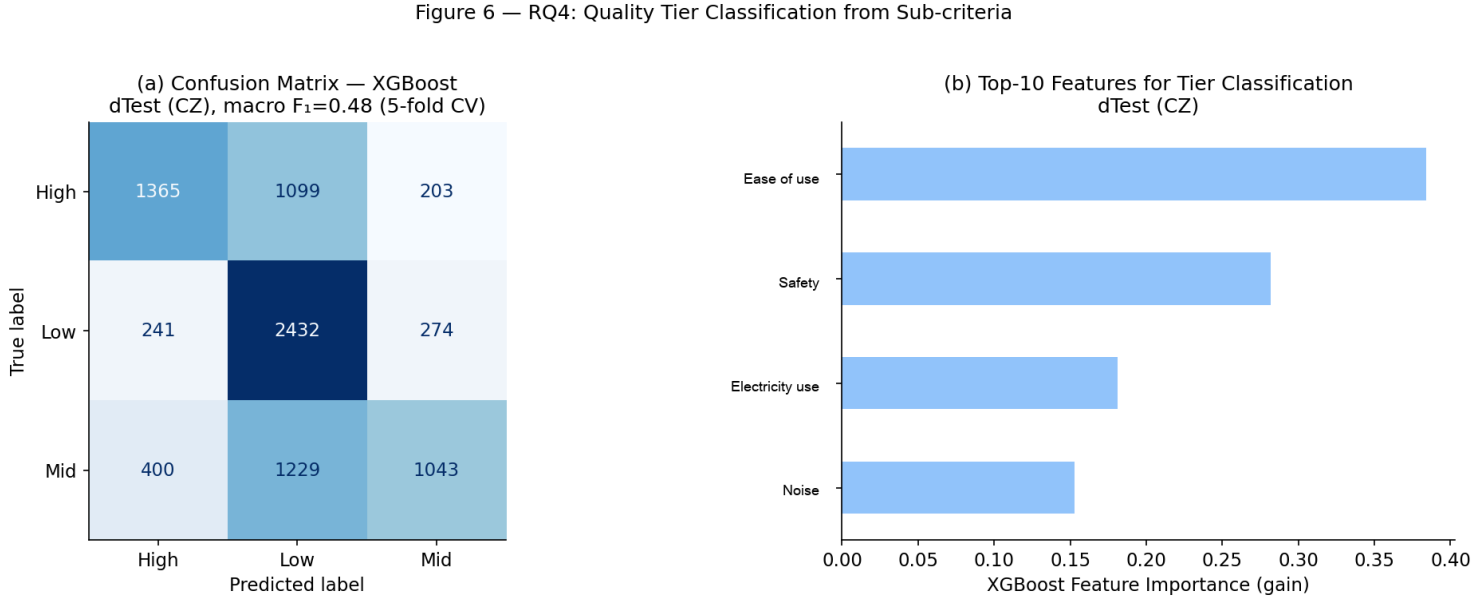


Figure 5

RQ4 — (a) Confusion matrix for XGBoost quality tier classification on dTest (macro-F_1 = 0.48, 5-fold CV). Most errors occur at adjacent tier boundaries. (b) Top-10 XGBoost feature importances.