International Journal of Hospitality Management 111 (2023) 103493

Contents lists available at ScienceDirect

International Journal of Hospitality Management

journal homepage: www.elsevier.com/locate/ijhm

'Apparent' and actual hotel scores under Booking.com new reviewing system

Veronica Leoni [a,b,*,1], David Boto-García [c,2]

[a] Center for Advanced Studies in Tourism, University of Bologna, Italy
[b] Department of Applied Economics, University of the Balearic Islands, Spain
[c] Department of Economics, University of Oviedo, Spain

ARTICLE INFO

Keywords:
Booking.com
Hotel
Rating scores
Scale effects
Propensity score matching

ABSTRACT

In September 2019, Booking.com changed its reviewing system based on the simple average of six items on a 2.5–10 scale by an unrestricted valuation on a 1–10 scale. This change has resulted in the drop of *observed* average scores. However, it is unclear which part of the shrinkage is due to the scale adjustment and which to priorly neglected aspects that consumers consider when valuing their satisfaction with the hotel stay. Using a dataset of more than 429,000 individual reviews for hotels in Madrid, Barcelona, Rome, Milan and Lisbon before and after the change, this paper disentangles *apparent* from *actual* changes in scores produced by the new scoring system. Using linear regressions and Propensity Score Matching, we show that, once the scale effect is left out, the new system has led to an increase by around 0.1 points in the *actual* valuation. Our results are potentially explained by the existence of unpacking effects.

1. Introduction

The hospitality literature has traditionally devoted great attention to guests' evaluation of consumer services because of its important economic and managerial implications. Together with the well-known positive effects of satisfaction on repeat purchases (Choi and Chu, 2001), intention to recommend (Akinci and Aksoy, 2019) or reputation building (Hörner, 2002), the development of Internet user-generated opinions in websites like TripAdvisor.com or Booking.com has increased the visibility of good and bad experiences to others. Since the quality of accommodation stays is uncertain *ex ante*, prospective consumers use online reviews and ratings as a source of information when making their accommodation choices (Fernández-Barcala et al., 2010; Martín-Fuentes, 2016). That is, opinions left by other consumers on online platforms act as quality cues. In this regard, electronic word-of-mouth (eWOW) has been shown to increase hotel consideration (Vermeulen and Seegers, 2009), booking intentions (Casaló et al., 2015), consumer trust (Sparks and Browning, 2011), hotel occupancy rates (Viglia et al., 2016) and brand meaning cocreation (Borges-Tiago et al., 2021). As a result, online review scores are a strong predictor of market survivability (Naumzik et al., 2022) and financial profitability and sales (Anagnostopoulou et al., 2020; Nieto-Garcia et al., 2019).

Given the importance of user-generated hotel scores, several scholars have studied the content and characteristics of hotel reviews in online platforms like TripAdvisor (e.g., Taecharungroj and Mathayomchan, 2019) and Booking.com (e.g., Mariani et al., 2019). The latter is nowadays the leading Online Travel Agency (OTA) (Martín-Fuentes and Mellinas, 2018), counting more than 28 million accommodations worldwide. Most studies highlight that the distribution of hotel scores is highly left skewed, with most hotels exhibiting scores over 8 and very few with bad ratings. In the case of Booking.com, Mellinas et al. (2015) show that this is because the reviewing system implemented by the platform uses a 2.5–10 scale for which the mid-point is 6.25. Since consumers generally judge scales as if they were defined on the 1–10 interval (Preston and Colman, 2000), a hotel score of 7 might be interpreted as a medium-high score whereas it actually should be deemed simply as good. In this respect, existing works that have compared hotel scores in Booking.com with other online platforms like Priceline (Mellinas et al., 2015), TripAdvisor (Martín-Fuentes et al., 2018) and Agoda and Atrapalo (Martin-Fuentes et al., 2020) show that average ratings for the same hotels are higher in Booking.com. Accordingly, hotel scores in Booking.com could be inflated, making it difficult for an average

* Correspondence to: Center for Advanced Studies in Tourism, Via Angherá, 22, 47921 Rimini, Italy.
E-mail address: veronica.leoni3@unibo.it (V. Leoni).
[1] ORCID ID: 0000-0001-8419-4241
[2] ORCID: 0000-0001-8065-0983

https://doi.org/10.1016/j.ijhm.2023.103493
Received 28 July 2022; Received in revised form 27 March 2023; Accepted 1 April 2023
Available online 12 April 2023

internet user to properly assess the quality of a hotel.

In September 2019, Booking.com changed its reviewing system by shifting from computing hotel overall scores as the average of scores assigned to six aspects on a 2.5–10 scale to asking consumers to directly rate their hotel experience on a 1–10 scale. Moreover, while in the old system consumers were asked to choose among four smileys, nowadays they are required to select an integer score on a slider. This change from a multi-dimensional to a single-dimensional reviewing system leads to two main effects on observed rates. On the one hand, by construction, the change in scale produces a drop in overall scores because the midpoint is now 5 rather than 6.5. On the other hand, enabling users to directly enter their overall score makes it to likely depart from the average of scores given to the six traditional aspects, something which is theoretically sustained by 'unpacking effects' (Angelini et al., 2017; Van Boven and Epley, 2003), dimensional rating bias (Schneider et al., 2021) and priming effects (Chen et al., 2018). This issue has raised concerns among both hotel managers and prospective users who are not able to figure out the sources of the score discrepancy (Mellinas et al., 2021). On these grounds, the following research question emerge: is the observed drop in scores a mere scaling issue or does it reflect a truly lower satisfaction through previously neglected dimensions and/or behavioural biases like *unpacking effects*?

The goal of this paper is to study the effects of the change in Booking.com review system on both '*apparent*' hotel scores (without scale adjustment) and '*actual*' scores (once the scale has been adjusted). The latter captures score differences due to the following facts: (i) consumers might internally attach different subjective weights to the different hotel quality dimensions (instead of forcing all of them to enter the overall valuation with equal weights); (ii) other hotel quality dimensions beyond the fixed six aspects defined in the old system might be considered now, and (iii) consumers can offer a more accurate valuation of their stay through the wider range of the available rating scale (Preston and Colman, 2000).

We use a large dataset involving 429,304 individual reviews in Booking.com from five of the most important Southern European cities (Madrid, Barcelona, Rome, Milan and Lisbon). The dataset covers the period October 2018-January 2019 (old scoring system) and October 2019-January 2020 (new scoring system). As opposed to previous descriptive analyses (Amblee and Ullah, 2022; Mellinas and Martin-Fuentes, 2021), we implement Propensity Score Matching (PSM) technique to match individual review scores for the same hotel and for travelers with similar characteristics who rate their accommodation before and after the change. Therefore, our econometric modelling strategy allows us to estimate the effect of the change in Booking.com reviewing system on hotel scores.

Our results show that *observed* hotel scores decreased, on average, by around 0.15 points after the implementation of the new system. However, once scale differences are removed, we find that actual scores increased by about 0.1 points on average. This effect is quite homogeneous across cities and remains robust under a battery of checks. We discuss the sources and implications of these differences in detail later in the paper.

This work makes three contributions to the tourism and hospitality literature. First, evidence on how the reviewing change in Booking.com has affected observed and actual hotel scores is scarce to date. Some previous studies have predicted that the new Booking.com review system might cause observed changes in hotel scores that dot not truly reflect changes in customer satisfaction (Mellinas and Martin-Fuentes, 2021) or studied how to harmonize overall scores that mix reviews from the old and the new system (Amblee and Ullah, 2022). Kim et al. (2022) have gone beyond and examined the causal impact of the new system on overall scores using a difference-in-differences (hereafter DiD) research design. Whereas these studies focus on changes in overall scores at the hotel level, we still know little about how review systems' design affect consumers' *individual* scores in online platforms. To the best of our knowledge, this is the first paper that attempts to disentangle *scaling* from *true* satisfaction change in hotel scores at the review level. Our work therefore sheds light on the important implications of changing crowd evaluation mechanisms and scale ranges in the hospitality industry. From this viewpoint, the paper adds to a large literature on the non-neutral implications of choice architecture and defaults (Cronqvist et al., 2018; Thaler, 2018).

Second, we offer a theoretical decomposition of the observed gap in hotel scores before and after the change. In particular, the score is decomposed into four parameters that we label as the *weighting*, the *scaling*, the *variability* and the *omitted satisfaction* components. Our third contribution is methodological. We apply PSM based on travelers' characteristics (length of stay, travel party composition and nationality) with exact matching of reviews made for the same hotel, before and after the change. As such, our estimates can be given a causal interpretation (Heckman et al., 1997, 1998). Inverse Probability Weighting regression adjustment (Hirano et al., 2003) is also used as a robustness check. As compared to DiD methods, our econometric strategy avoids potential comparability problems when working with individual reviews associated with consumers' self-selection into OTAs (Chen et al., 2021). Moreover, we use a large dataset covering five of the most important tourist cities in Southern Europe. In this regard, there is an increasing awareness among scholars in hospitality and tourism about the advantages and possibilities of using big data analytics (Lyu et al., 2022; Mariani, 2020). Our results offer the great advantage they are not sensitive to the specific idiosyncrasies of a city or to cultural factors, which ensures a high external validity to our findings.

2. Literature review

2.1. Accommodation reviews in used-generated-content platforms

A large body of research has examined consumer reviews and ratings in different user-generated-content platforms. Detailed reviews of on the topic can be found in Leung et al. (2013), Serra Cantallops and Salvi (2014) and Magnani (2020). Some stylized facts from this literature are the following. First, scores distributions are highly left skewed, with mean ratings around 8 (Mariani and Borghi, 2018; Mariani et al., 2019; Racherla et al., 2013). Some authors indicate that hotel scores in online platforms might not be accurate and credible because of manipulation from the firms' perspective (Mayzlin et al., 2014) together with consumers' fake reviews (Luca and Zervas, 2016) and self-selection into online platforms (Shen et al., 2021). Nonetheless, Martín-Fuentes (2016) provides evidence that scores in websites like TripAdvisor or Booking are well-aligned with official hotel star-rate classifications. Furthermore, Smironva et al. (2020) show that both online and offline rating show similar distributions, supporting the notion that valuations offered in online settings are similar to the ones provided offline.

Second, the presence of hotels in user-generated online platforms depends on the size, category and independent vs. chained nature of the hotel (Martín-Fuentes and Mellinas, 2018). In this vein, autonomous hotels are better in satisfying the customer in value for money while those affiliated to small chains have a greater ability to satisfy in comfort (Moreno-Perdigón et al., 2021). Third, the submission device and the presentation format for the reviews seems to be non-neutral. For instance, Mariani et al. (2019) show that mobile reviews are shorter but with greater valence than that from desktop eWOM. Kim et al. (2021) provide evidence that visual information in the form of photos of rooms and facilities exert a positive effect on review ratings and sentiments. Fourth, attribute ratings are mutually dependent, implying that the assessment of satisfaction with one hotel dimension is affected by others (Nicolau et al., 2020a). Accordingly, dissatisfaction with one attribute has the power to drop the overall evaluation through consumers' biased tendency to punish the hotel if one dimension is not as desired (Mellinas et al., 2019).

Fifth, online ratings vary across tourist profiles. Online reviews differ depending on the length of the stay (Kim and Han, 2022), nationality

(Rita et al., 2022) or the composition of the travel party (Radojevic et al., 2015; Banerjee and Chua, 2016). Furthermore, factors like customer's cultural traits (Mariani and Predvoditeleva, 2019), the use of domestic vs. foreign language (Mariani et al., 2019), geographic and psychic distance (Phillips et al., 2020) or traveller experience (Gao et al., 2018) have been shown to shift hotel ratings. In this respect, recent evidence by Nicolau et al. (2022) shows that the use of language by the hotel staff exerts positive but asymmetric effects on satisfaction scores. Additionally, ratings notably differ depending on whether the monetary component is considered. For example, Nicolau et al. (2020b) indicate that hotel scores are lower when evaluating value for money as compared to satisfaction scores that only involve non-monetary aspects.

Lastly, another typical finding is that consumers seem to be sensitive to prior ratings when evaluating their stay at a hotel, particularly when those ratings are above the average (Cicognani et al., 2021). This pattern is consistent also with herding and informational cascades documented in other online rating systems like those of movies (Lee et al., 2015) or restaurant dishes (Cai et al., 2009). Furthermore, in line with the theory of first-impression bias developed by Rabin and Schrag (1999), consumers who choose a hotel due to recommendation have been found to be more subsequently satisfied with hotel services because people tend to sustain *ex-ante* perceptions (Boto-García et al., 2021).

2.2. Choice architecture and measurement scales

In recent years, a large body of research has started to pay attention to the role of choice architecture on people's behaviour, defined as the construction and definition of the set of options available to decision makers. This stream of literature has shown that default options can act as nudge-like interventions aimed at framing and targeting several choices, some of which are deemed better from a social welfare viewpoint (Cronqvist et al., 2018; Thaler, 2018). The way choice options are defined can exert non-negligible impacts on how people choose and behave. For instance, in a famous study, Thaler and Benartzi (2004) provide evidence that reframing the "decision to save" from reducing consumption now to how much future increase in salary will be allocated to savings reduces the impact of impatience and increases savings. Using a series of experiments, Angelini et al. (2017) show that asking people to report satisfaction with life domains before overall satisfaction with life generates 'unpacking effects' that shift upwards the subsequent mean overall life satisfaction evaluations. Based on two quasi-experiments in Britain, Conti and Pudney (2011) exploit a change in question design to show that apparently minor differences in survey design can produce misleading conclusions.

In the tourism context, several scholars have shown that the definition of measurement scales and choice alternatives can impact the obtained results (Dolnicar and Grün, 2013; Araña and León, 2013). The definition of the number of alternatives in Likert-type measures when assessing satisfaction has also been shown to matter for reliability and response bias (Pizam et al., 2016). The study by Preston and Colman (2000) shows that scales with response categories above 7 are better suited for the purpose of assessing consumer satisfaction in terms of reliability, validity and discriminating power. They document that people prefer 10-point scales. Based on a between-subject experiment, Chen (2017) reports that five-star rating systems can make users feel more fun than binary-textual and visual rating systems. In the light of all this evidence, how online platforms define hotel valuation metrics is likely to be non-neutral and deserves a closer examination. In the following section, we review existing differences in hotel scores depending on the platform, with particular attention to Booking.com's review system design.

2.3. Booking.com review system

Booking.com and TripAdvisor are the leading platforms for electronic word-of-mouth communication in the hospitality industry. The former is said to be more reliable because whereas reviews on TripAdvisor are not constrained to those who stayed overnight at the accommodation (anyone can post a comment), Booking.com implements a verified user system by which only actual consumers can rate the hotel. In this regard, consumers have been shown to be quite sceptical about eWOW because of the users' anonymity (Sparks and Browning, 2011; Leung et al., 2013). Martín-Fuentes et al. (2018) compare hotels reviews in both platforms and document that although scores and rankings in the two platforms are positively correlated, the score means are quite different. Similarly, Figini et al. (2020) analyse the distribution of scores of the two leading platforms. These authors find that scores left on non-verified systems are more volatile, also displaying a slower convergence to longer-term scores.

Most of the literature has assumed that Booking.com uses a 10- or 11-point Likert scale, as it is customary in satisfaction studies. However, Mellinas et al. (2015) were among the first that warned that Booking.com actually used a scoring system on a 2.5–10 scale. The fact that the lower bound of the scale was 2.5 rather than 0 or 1 explains the common finding that most hotels exhibit scores above 7. Martín-Fuentes et al. (2018) argue that the higher scores of Booking.com compared to TripAdvisor are in fact the result of the differences in the scales used.

Tourism scholars have investigated the implications of the differences in the scale used for ratings. Mariani and Borghi (2018) exploit more than 1.2 million online reviews for hotels in London and show the overall distribution of scores is left-skewed, with most hotel scores concentrated in the 7.5–10 interval. Interestingly, skewness is associated with hotel class. These authors argue that the non-normality of ratings' distribution is highly dependent on the 2.5–10 scoring system deployed by Booking.com. Mellinas et al. (2015) compare the scores in Booking.com with those in Priceline (which uses a 1–10 scale) for the same hotel. Average scores are found to be significantly higher in Booking.com, being the difference larger for hotels with low scores. Similarly, Martín-Fuentes et al. (2020) compare hotel scores in Booking.com with several other user-generated content platforms and obtain some puzzling results: as expected, Booking.com and Agoda (which uses a 2–10 scale) give higher rating scores than other OTAs like Atrapalo, Travel Republic and hotel reservation services (where 0–10 and 1–10 scales are in use); however, when scores are rescaled, Booking.com and Agoda produce the worst ratings. Bigné et al. (2020) do not detect significant differences in online review scores across 11 different platforms. Interestingly, these authors document some inconsistencies between scores received by attributes individually and the overall score.

In a recent study, Kim et al. (2022) evaluate the change in overall hotel scores in Booking.com caused by the new scoring system. The authors use reviews in Priceline in the same period as the comparison group. Using a difference-in-differences design, they find that the new system has a positive impact on highly positive reviews but a negative one on those with medium-to-high scores. Their analysis focuses on guests' evaluation at the hotel level, with an emphasis on the change in the scores' distribution. We, instead, take individual scores as the unit of analysis, preventing the mix between reviews under both systems. In doing so, we focus on changes in mean values by a vis-à-vis matching of travellers of the same profile before and after.

3. Conceptual framework

After the hotel stay, customers receive an e-mail from Booking.com asking for their opinion about the experience in the hotel. They are required to grade six different aspects using four different smiley faces that have their numeric equivalent (Poor=2.5; Fair=5; Good=7.5; Excellent=10). The six aspects considered are the following: cleanness, comfort, location, facilities, staff and value for money. A visual representation of this traditional multi-evaluation system is presented in Fig. 1. In the former reviewing system, Booking.com calculated the overall score as the plane average of the ratings assigned to each of the six aspects. Since the minimum valuation allowed is 2.5, the overall

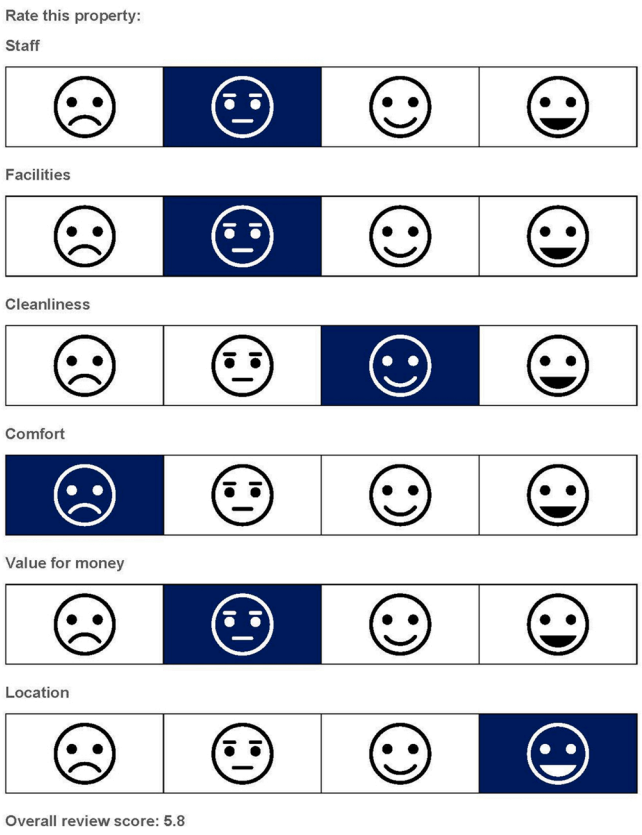

Fig. 1. Example of the review process of the old Booking.com system. Source: own elaboration based on Booking.com multi-attribute review system.

score was also bounded between 2.5 and 10.

In September 2019, Booking.com introduced two important changes in the scoring system: First, they switched the 2.5–10 scale to a more conventional 1–10 scale. Second, consumers are now allowed to select an overall score themselves. This implies that the visible score associated to each hotel is no longer the plain average of the six aspects considered (whose valuation has been nonetheless kept in the 2.5–10 scale and measured through smileys but does not enter in the overall score evaluation). Fig. 2 presents an illustration of the new review system. As a result of these changes, hotel scores have shrunk, especially among hotels with current low and medium scores (Mellinas and Martin-Fuentes, 2021). Assuming that the quality of hotel services is the same before and after the change (i.e., the underlying quality has not been affected by the valuation change), the new system makes scores to be individually lower because the minimum value allowed is now lower. Since Booking.com uses a 3-year rolling window, the new individual scores also pull down aggregate scores. As discussed in Mellinas and Martin-Fuentes, 2021, some hoteliers and prospective guests might be concerned when they notice overall scores have dropped by almost one

Fig. 2. Example of new Booking.com review process. Note: The evaluation process displayed in Figure 1 is still part of the new system, but the new overall score is independent of it. As shown in this example, the overall score is now 8.0, which does not correspond to the plain average of the six sub-items (5.8). Source: own elaboration based onBooking.com review system since September 2019

point. This might exert non-negligible negative effects on future booking intentions through the above-mentioned herding-behaviour mechanism (Figini et al., 2020).

In a seminal study, Schneider et al. (2021) provide a theoretical characterization of the expected differences in service evaluations between multi-dimensional (the review poster is required to make several ratings) and single-dimensional scoring systems (only an overall rating is required). These authors show that although multi-dimensional ratings tend toward the mean of overall ratings, the choice among one type or another is non-neutral. In particular, single-dimensional systems tend to prime the most relevant dimension of the service experience whereas multi-dimensional systems, apart from imposing greater cognitive burden and survey fatigue, make some dimensions more accessible to memory. A key insight from their work is that overall ratings do not necessarily match the mean of multidimensional ratings; they could be either higher or lower. Their framework closely relates to the 'unpacking effect' firstly developed by Rottenstreich and Tversky (1997) and later presented in Van Boven and Epley (2003): the whole is less than the sum of its parts, making overall judgements to depart from the sum of individual valuations. At the empirical level, both positive and negative deviations between overall (single-dimensional) and item-specific (multi-dimensional) valuations have been widely documented (Angelini et al., 2017; Bateman et al., 1997; Chen et al., 2018).

In what follows, we provide a theoretical characterization of the expected effects of Booking.com review system change on hotel scores and why the gap in scores could be either positive or negative. Let us assume the overall latent satisfaction of a given consumer i with hotel j services at period t (SAT_{ij}^{*}) can be expressed as an additive sum of three terms as follows:

$$SAT_{ij}^{*} = \frac{\sum_{k=1}^{6} w_{ijk} SAT_{ijk}}{\sum_{k=1}^{6} w_{ijk}} + \eta_{ij} + \varepsilon_i \quad (1)$$

where SAT_{ijk} measures the satisfaction with aspect k for $k = \{cleanness, comfort, location, facilities, staff, valueformoney\}$, w_{ik} indicates the individual weight assigned to each aspect in the overall valuation, η_{ij} refers to the satisfaction with any hotel characteristic other than the six aspects considered (e.g. sleep or food quality), and ε_i is a zero-mean normally distributed error term that reflects individual idiosyncratic effects that affect overall valuation (e.g., mood). The term η_{ij} is likely to be heterogeneous across consumers and hotels so that $\eta_{ij} \sim N(\mu, \sigma_\eta)$.

Before the change in the scoring system (hereafter denoted by subindex 0), we observe in contrast the following indicator of satisfaction:

$$SAT_{ij_0} = \frac{\sum_{k=1}^{6} SAT_{ijk}}{6} \quad (2)$$

That is, under the old system, hotel scores for each guest are calculated as the average satisfaction with the six aspects under the restrictive assumption that $w_{ik} = 1 \forall k$. Beyond that, there could be measurement error in SAT_{ijk} because hotel guests are restricted to choose among four discrete rates (2.5, 5, 7.5 and 10) for each aspect k. Therefore, hotel scores in the old system might be a biased representation of guests' underlying satisfaction.

One caveat is in order. The individual assessment of satisfaction with *each aspect k* could be *also* affected by idiosyncratic factors ε_{ik}so that $SAT_{ijk} = SAT_{ijk}^{*} + \varepsilon_{ik}$, where SAT_{ijk} indicates the observed rating assigned to item k and SAT_{ijk}^{*} denotes the latent satisfaction with item k. For simplicity, we assume that the expected value of the idiosyncratic factor is equal to zero ($E(\varepsilon_{ijk}) = 0$). Under this assumption, we therefore consider that positive and negative deviations from the real satisfaction due other factors like weather (Štumpf et al., 2022) or mood effects (Kim and Mattila, 2010) are compensated in the population.

After the review system change (hereafter denoted by subindex 1), the consumer has the freedom to enter any integer value on a 1–10 scale. As a result, the overall score we observe can be expressed as follows:

$$SAT_{ij_1} = \frac{\sum_{k=1}^{6} w_{ijk} SAT_{ijk}}{\sum_{k=1}^{6} w_{ijk}} + \eta_{ij} + \varepsilon_{ij} \quad (3)$$

where SAT_{ij_1} gathers a weighted average of satisfaction with the six aspects ($\frac{\sum_{k=1}^{6} w_{ijk} SAT_{ijk}}{\sum_{k=1}^{6} w_{ijk}}$), satisfaction with other facets (η_{ij}) and idiosyncratic effects (ε_{ij}). As such, the score provided under the new system is likely to be a more accurate measure of the real overall satisfaction with the hotel so that SAT_{ij_1} is an unbiased measure of SAT_{ij}^{*}.

If we now take the expectation of the difference in the observed scores between before and after the change, we get the following:

$$Diff_{ij} = E(SAT_{ij_1} - SAT_{ij_0}) = E(\frac{\sum_{k=1}^{6} w_{ijk} SAT_{ijk}}{\sum_{k=1}^{6} w_{ijk}} - \frac{\sum_{k=1}^{6} SAT_{ijk}}{6} + \eta_{ij} + \varepsilon_{ij}) \quad (4)$$

Since it holds that $E(\varepsilon_{ij}) = 0$ and $E(\eta_{ij}) = \mu$, the difference in means can be rewritten as the sum of four components as follows:

$$Diff_{ij} = \gamma + \delta + \theta + \mu \quad (5)$$

The term γ captures the part of the difference that is due to imposing equal weights for the six aspects in the old system as opposed to allowing the consumer to freely weigh them. Several works have shown that consumers do not assign the same importance to all facets that integrate a hotel stay (Nieto-Garcia et al., 2019; Zhang et al., 2021). We call γ the *weighting* component. This term can take any value since $\frac{\sum_{k=1}^{6} w_{ijk} SAT_{ijk}}{\sum_{k=1}^{6} w_{ijk}} - \frac{\sum_{k=1}^{6} SAT_{ijk}}{6} \geq 0$ or vice versa. The parameter δ measures the part of the difference that emerges due to scale differences: recall that whereas SAT_{ij_1} is bounded between 1 and 10, SAT_{ij_0} only ranges between 2.5 and 10. As such, $\delta < 0$ by construction. We call δ the *scaling* component. The term θ reflects the part of the difference that is due to the reduced variability in SAT_{ijk} in the old system (people can choose among only four discrete values: 2.5, 5.0., 7.5 or 10.0) as compared to SAT_{ijk} in the new system (which can take any continuous value on the 1–10 interval). In principle, θ could be positive or negative because the measurement error associated with SAT_{ijk} can be of either sign. We label θ as the *variability* component. Finally, the parameter μ gathers the mean difference in scores attributed to the potential consideration of more dimensions when judging hotel services under the new system. In this regard, some authors have shown that hotel guests value aspects like sleep and food quality (Zhang et al., 2021) or WiFi flaws (Mellinas and Nicolau, 2020) when assessing their overall satisfaction. We call $\mu$the *omitted satisfaction* component.

The value of $Diff_{ij}$ would inform about the change in 'apparent' scores, which are the ones prospective guests see in the platform when they are searching for an accommodation. Since part of the discrepancy is attributed to the *scaling* component, we could remove that part of the score gap by re-scaling SAT_{ij_0} on the 1–10 interval as follows:

$$\widetilde{SAT}_{ij_0} = 10 * \frac{\left(\frac{\sum_{k=1}^{6} SAT_{ik}}{6} - 2.5\right)}{7.5} \quad (6)$$

If we take again the difference in expectations, we have:

$$\widetilde{Diff}_{ij} = E(SAT_{ij_1} - \widetilde{SAT}_{ij_0}) = \gamma + \theta + \mu \quad (7)$$

where now $\widetilde{Diff}_{ij}$ measures the difference in 'actual' scores leaving scale effects (δ) aside. Because γ, θ and μ can take any value, $\widetilde{Diff}_{ij}$ can be either positive or negative, in line with Schneider et al. (2021).

4. Data and methods

4.1. Dataset and descriptive statistics

The data used in the analysis have been retrieved from Booking.com website using an ad-hoc designed web-crawler. The crawler allows us to obtain information about reviews both at the hotel and at the individual level. Since we are interested in potential changes in individual ratings, we work with the review-level dataset. Since the new system was implemented in September 2019, and to avoid overlapping with the changes in demand induced by COVID-19 pandemic (starting on March 2020), we use data for the first four months under the new system (October 2019-January 2020). For comparability purposes, we also consider reviews under the old system for the same period one year ahead (October 2018-January 2019). This is done to ensure a meaningful comparison leaving aside potential seasonal differences. Nonetheless, a wider comparison in terms of the time period window (October 2018-August 2019) is also used for robustness (see subsection 5.2).

The dataset involves 429,304 individual reviews from guests who stayed overnight in 1458 hotels located in five of the most relevant South-European destinations: Milan (252), Rome (639), Madrid (77), Barcelona (192) and Lisbon (239). Each observation consists of the overall individual score left by tourist *i,* staying in hotel *j,* at period *t* (SAT_{ijt}). The period corresponds to the exact calendar date when the review was left on the platform, which by definition is subsequent to the hotel stay. In addition to the city, the hotel and the date of the review, we also avail of the following information about each guest's profile: travel party composition (solo traveller, travelling in a couple, with the family or in a group), guests' nationality and the length of the stay at the hotel (number of nights).

Fig. 3 presents histogram plots of observed hotel scores at the city level, both before (SAT_{ijt_0}) and after (SAT_{ijt_1}) the review system change. Overall, scores exhibit the classic left-skewed distribution documented in previous studies (Mariani and Borghi, 2019). Considering the pooled dataset, the average score is 8.70 under the old system but 8.52 under the new one. Note that, as discussed in Section 3.1, by construction SAT_{ijt_1} is expected to have a lower mean.

Fig. 3. (cont.) Histogram of scores per city before (Pre) and after (Post) the review system change.

Table 1 presents descriptive statistics of the variables used in the analysis, separately for each city. The share of observations under the new system is about 63%, which allows for a fair comparison between treated and controls. From a geographical standpoint, observations are quite homogenously distributed across cities, with 25% of them belonging to hotels located in Rome, 22% in Lisbon, 13% in Milan, 17% in Madrid and 27% in Barcelona. On average, each hotel received 2309 scores, on average, during the study period but with high heterogeneity across hotels (SD=2286). Regarding the type of traveller, most reviewers travelled in couple (41%), followed by family (20%), solo travellers (22%) and groups (17%). About 20% of reviews belongs to domestic travellers and, on average, tourists stayed 2.6 days at the hotel.

4.2. Methodology

Our empirical strategy aims at isolating the effect of the new reviewing system by comparing similar reviews left before and after the change. For this purpose, we first rely on standard regression analysis. Next, we move to formal causal inference methods.

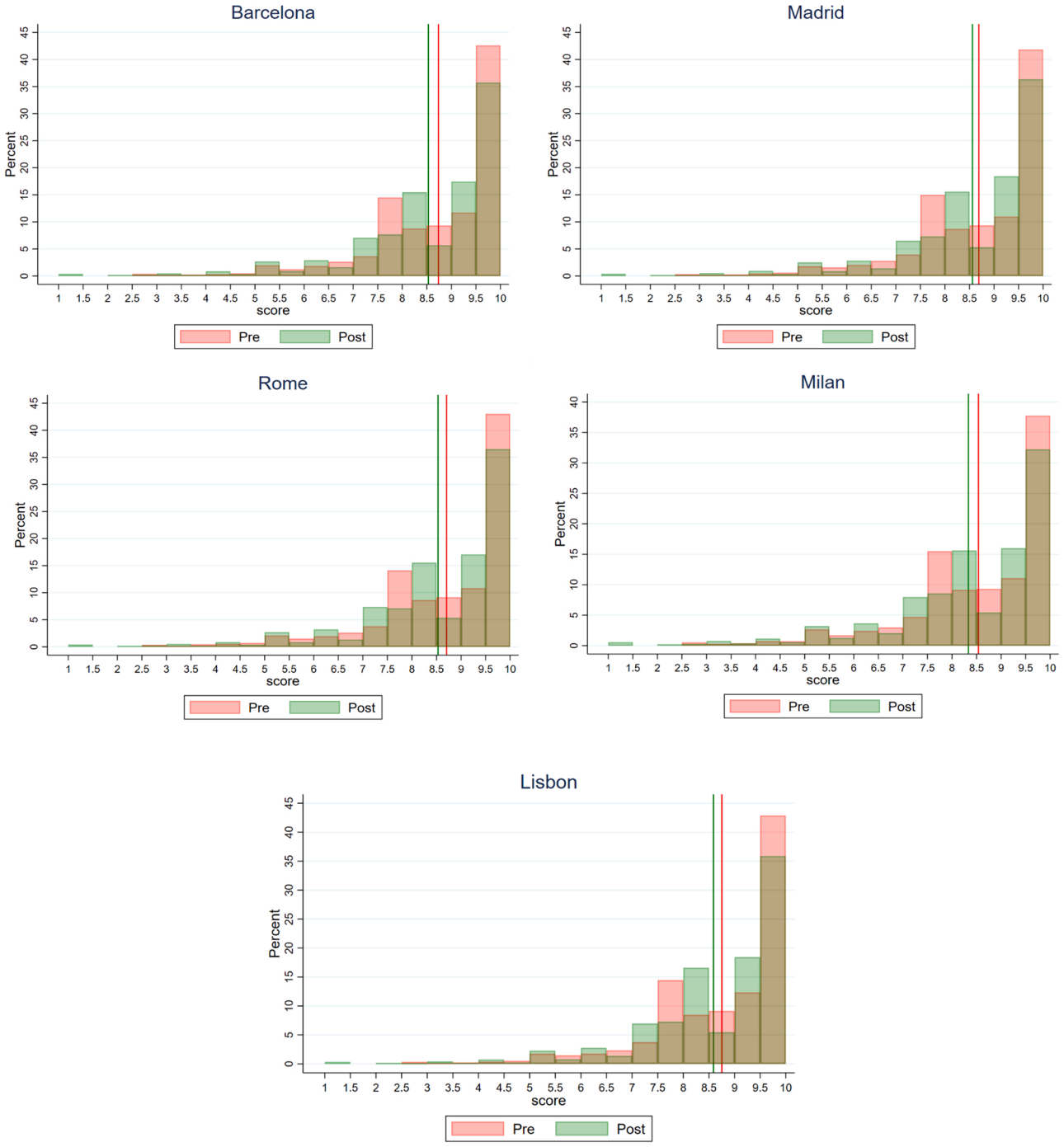

Fig. 3. Histogram of scores per city before (Pre) and after (Post) the review system change.

4.3. Regression analysis

We first propose the following regression model:

$$SAT_{ijt} = \alpha + \tau NewSystem_{ijt} + \beta X_{ijt} + HotelFE_j + M_t + \varepsilon_{ijt} \quad (8)$$

where $Score_{ij}$ is the individual score left by individual i who stayed at the hotel j in period t; X_{ijt} is a vector of tourist characteristics (nationality, travel party indicators and length of stay); $Hotel\ FE_j$is a set of hotel fixed effects absorbing variation in satisfaction stemming from hotel generic attributes; M_t are month fixed effects; τ is the main parameter of interest, which captures the effect of the change in the reviewing system ceteris paribus; and ε_{ij} is a normally distributed error term.

The model in (8) is estimated using Ordinary least Squares both using the original scores (SAT_{ijt}) as well as its rescaled version ($\widetilde{SAT}_{ijt}$) as dependent variables (i) for each city separately, and (ii) for the pooled dataset. Tobit regression is used as a robustness check (see subsection 5.2). It is important to make it clear that we assume cardinality in hotel satisfaction because overall scores are continuous indicators (at least in the old system for SAT_{ijt} and for all observations in the rescaled version $\widetilde{SAT}_{ijt}$).

4.4. Matching estimator

As an alternative estimation strategy, we move to matching estima-

Table 1
Descriptive statistics of the city sub-samples and the pooled sample.

	Barcelona		Madrid		Rome		Milan		Lisbon		Pooled sample	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
SAT	8.61	1.49	8.61	1.49	8.59	1.53	8.41	1.61	8.65	1.43	8.59	1.50
SAT rescaled	8.51	1.59	8.51	1.59	8.50	1.63	8.30	1.72	8.55	1.53	8.49	1.61
Length of Stay	2.83	1.58	2.34	1.48	2.65	1.65	2.00	1.24	2.55	1.69	2.55	1.59
	%		%		%		%		%			
New system	60.37		62.46		62.90		61.24		61.66		61.67	
Single	22.14		21.61		20.41		18.56		19.57		20.61	
Couple	36.02		42.79		45.34		34.53		46.39		41.27	
Group	20.79		16.13		14.06		18.56		14.62		14.41	
Family	21.03		19.45		20.17		20.94		19.40		20.25	
Domestic	12.74		43.68		27.04		30.52		17.08		19.95	
October	27.00		29.84		28.49		20.09		26.31		27.86	
November	32.43		29.79		32.19		28.94		29.60		30.70	
December	17.73		23.68		22.75		24.96		25.72		24.09	
January	22.14		16.67		16.54		16.99		18.34		17.33	
N	117,305		53,150		109,574		56,913		92,362		429,304	

tors within the potential outcomes framework originally developed by Rubin (1974). Even though the regression model in (8) cleans up the part of the variability in scores associated with time-invariant hotel attributes and tourist personal characteristics, τ could be an imprecise estimate of the causal effect of the new review system on hotel scores in case there is an unbalance in the characteristics of hotel guests before and after the change (limited overlap). To guarantee a better comparison, we match reviews under the new system with reviews under the old system based on the similarity of their propensity scores $p(X_{ijt})$, defined as the conditional probabilities of being in the new system given the tourist-specific covariates (Rosenbaum and Rubin, 1983):

$$p(X_{ijt})=Prob(NewSystem_{ijt}=1|X_{ijt}) \tag{9}$$

The propensity scores are estimated using a Logit regression and units are subsequently matched using kernel matching (Heckman et al., 1998). We look for statistical 'twins' of individual reviews in the old system based on similarity in the guest profile to derive the counterfactual. Importantly, we apply exact matching by hotel, meaning that only reviews belonging to the same hotel are compared. This is done using the *kmatch* Stata package with automatic bandwidth selection (Jann, 2017).

Matching on a single propensity score rather than on a set of covariates helps to alleviate the dimensionality problem to a single scalar dimension (Rosenbaum and Rubin, 1983), as well as ensuring *balancing* and *unconfoundedness*. The first implies that treatment and covariates are independent conditional on the propensity scores. The latter implies that assignment to treatment is random conditional on the propensity-scores (Cerulli, 2015, pp.78–79).

Assuming that the 'treatment' ($NewSystem_{ijt}$) is exogenously given (i.e., the timing around the adoption of the new system is independent of tourist characteristics and their hotel experience), the Average Treatment Effect (ATE) of the new system on *apparent* satisfaction scores is given by averaging the pairwise differences in scores between matched treated and control reviews as follows:

$$ATE = E\{E[SAT_{ijt}|p(X_{ijt}), NewSystem = 1] - E[SAT_{ijt}|p(X_{ijt}), NewSystem = 0]\} \tag{10}$$

where as that for *actual* satisfaction scores is:

$$ATE = E\{E[\widetilde{SAT}_{ijt}|p(X_{ijt}), NewSystem = 1] - E[\widetilde{SAT}_{ijt}|p(X_{ijt}), NewSystem = 0]\} \tag{11}$$

The reader is referred to Heckman et al. (1998) and Cerulli (2015) for further methodological details.

Before moving on, it is important to indicate that we could use reviews from other OTAs as a comparison group and examine the change in Booking.com in a difference-in-differences (DiD) setup, as done by Kim et al. (2022). We disregard this approach in our setting for the following reason. When working at the hotel-level, the DiD method identifies the causal change in the overall score due to the system change by observing mean scores for the same hotel in a treated and non-treated platform. However, our goal here is different, as we aim to evaluate the change in reviews at the *individual* rather than aggregate level. When working at the individual level, reviews made by distinct consumers in different platforms might be difficult to compare. Besides platform selection effects by which consumers opt for gathering information in one OTA or another non-randomly (Chen et al., 2021), the platform design itself influences reviewers' sentiment and tone of voice. For instance, Rita et al. (2022) report that written reviews in TripAdvisor are much more positive than in Booking.com, being this strongly associated with guests' nationality. The structural difference of reviews on non-verified (such as TripAdvisor) and verified platforms (such as Booking.com) is also confirmed by Figini et al. (2020). Moreover, hotel-level overall scores mix reviews under the old and new system due to the three-year rolling window that Booking.com uses to construct the average score. Therefore, we prefer to create the counterfactual based on reviews for the same hotel, written by travellers sharing a similar profile, on the same platform, and during exactly the same period (October-January) one year before. This identification strategy allows also to leave aside potential seasonality effects.

5. Results

5.1. Main findings

Table 2 presents the coefficient estimates from separate OLS regressions per city. Results show a negative effect of the new system on the 'apparent' satisfaction, with an average reduction of about 0.15 points on individual scores. The magnitude of the effect is quite homogenous across cities (Barcelona= –0.147; Madrid= –0.147; Rome= –0.175 and Lisbon= –0.167), except for Milan where the drop appears to be stronger (–0.223 points). These results are in line with previous descriptive evidence on the effect of the change of the Booking.com reviewing system on hotel scores (Mellinas and Martin-Fuentes, 2021). However, when we move to the rescaled scores, we document an interesting pattern: *actual* satisfaction after the change is about 0.1 points higher. That is, leaving the scale effect aside, allowing consumers to freely rate their satisfaction in an unrestricted manner translates into a consistent average increase in satisfaction scores. Again, the pattern is very similar for all the cities (Barcelona=+0.109; Madrid=+0.118;

Table 2
Coefficient estimates from OLS regressions.

Sample	(1) Barcelona	(2) Barcelona	(3) Madrid	(4) Madrid	(5) Milan	(6) Milan	(7) Rome	(8) Rome	(9) Lisbon	(10) Lisbon
Explanatory variables	SAT	SAT rescaled	SAT	SAT rescaled	SAT	SAT rescaled	SAT	SAT rescaled	SAT	SAT rescaled
New system	-0.147 ***	0.109 ***	-0.147 ***	0.118 ***	-0.223 ***	0.072 ***	-0.176 ***	0.084 ***	-0.167 ***	0.085 ***
	(0.009)	(0.010)	(0.014)	(0.015)	(0.014)	(0.015)	(0.010)	(0.010)	(0.010)	(0.011)
Length of Stay	-0.021 ***	-0.023 ***	-0.021***	-0.021 ***	-0.001	-0.001	-0.024 ***	-0.026 * **	-0.016 ***	-0.018 ***
	(0.003)	(0.003)	(0.004)	(0.005)	(0.005)	(0.006)	(0.003)	(0.003)	(0.003)	(0.003)
Couple	0.054 ***	0.054 ***	0.124 ***	0.128 ***	0.027	0.027	-0.058 ***	-0.064 ***	0.001	-0.003
	(0.013)	(0.013)	(0.017)	(0.018)	(0.018)	(0.019)	(0.012)	(0.013)	(0.013)	(0.014)
Family	0.017	0.014	0.170 ***	0.178 ***	0.022	0.019	-0.051 ***	-0.057 * **	-0.026*	-0.027 *
	(0.014)	(0.015)	(0.020)	(0.022)	(0.020)	(0.021)	(0.014)	(0.015)	(0.015)	(0.016)
Group	0.076 ***	0.079 ***	0.137 ***	0.144 ***	0.043 **	0.046 **	-0.028 *	-0.030 *	0.036 **	0.036 **
	(0.013)	(0.014)	(0.021)	(0.022)	(0.020)	(0.021)	(0.015)	(0.016)	(0.016)	(0.017)
Domestic	-0.071 ***	-0.071 ***	-0.036 ***	-0.034 **	-0.010	-0.011	-0.059 ***	-0.058 ***	-0.091 ***	-0.094 ***
	(0.013)	(0.014)	(0.014)	(0.015)	(0.015)	(0.016)	(0.011)	(0.012)	(0.014)	(0.015)
Constant	8.433 ***	8.181 ***	7.274 ***	6.931 ***	8.149 ***	7.817 ***	9.692 ***	9.525 ***	9.802 ***	9.634 ***
	(0.056)	(0.060)	(0.046)	(0.049)	(0.210)	(0.225)	(0.284)	(0.303)	(0.061)	(0.065)
Hotel FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Month FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Number of hotels	192	192	78	78	264	264	684	684	240	240
Observations	117,305	117,305	53,150	53,150	56,913	56,913	109,574	109,574	92,362	92,362

Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Milan=+0.072; Rome=+0.084 and Lisbon=+0.085).

The regression results also indicate that hotel scores are negatively correlated with the length of the stay in all the cities except Madrid. That is, long stayers tend to give lower scores. Satisfaction with hotel services also varies depending on the composition of the travel party, in line with some previous evidence in the related literature (Radojevic et al., 2015; Banerjee and Chua, 2016). In this case, we detect relevant heterogeneity depending on the city considered. Couples and groups give higher rates than solo travellers in Barcelona and Madrid whereas the opposite pattern applies to Rome. Furthermore, domestic tourists are found to give significantly lower scores. This is consistent with Rita et al. (2022), who also report differences in online WOM depending on the tourist's nationality.

To ensure a more thorough comparison between scores before (non-treated) and after (treated) the change, in Table 3 we report the estimates of the Propensity Score Matching model (PSM), run on the five city subsamples and considering the pooled dataset. The corresponding coefficient estimates of the logit model are presented in Supplementary Material (Table A1). Cumulative distribution plots and box plots for raw and matched reviews per city are presented in Figures A1-A5. We impose the common support condition so that observations that do not have a close match based on characteristics are discarded, and that is why the matching procedure uses fewer observations per city. Results are consistent with the estimates in Table 2: after the introduction of the new scoring method, *apparent* satisfaction suffers a 0.162-point drop in Barcelona, 0.149 in Madrid, 0.221 in Milan, 0.155 in Rome and 0.155 in Lisbon. Pooling together all the cities, we find an Average Treatment Effect of –0.174. This effect is statistically significant in all cases. When looking at the effect on *actual* scores, PSM also produces very similar estimates to the ones obtained from the linear regression in Table 2: satisfaction increased on average by 0.096 points in Barcelona, 0.117 in Madrid, 0.075 in Milan, 0.084 in Rome and 0.099 in Lisbon. The overall average effect is estimated to be an increase in 0.09 points.

5.2. Robustness checks

We performed several robustness checks to inspect the sensitivity of our findings. First, we re-estimated the model in (8) using a Tobit regression. The results are presented in Table A2 in Supplementary Material. The estimates are very similar to the ones using OLS regression. Second, we run a pooled OLS regression using the review scores of the five cities together including interaction terms between *NewSystem* and the city dummies (Supplementary Material, Table A3) to formally evaluate potential heterogeneous effects by city. Relative to Barcelona, the drop in apparent scores is larger in Milan, Rome and Lisbon but smaller in Madrid. In contrast, the increase in actual scores is higher in Madrid followed by Barcelona, Rome, Lisbon and Milan. Third, the kernel matching used in Table 3 gives larger weights to controls with smaller distances (Epanechnikov kernel). Instead, we re-estimated the model using nearest-neighbour matching. The results are robust to this alternative matching procedure (Supplementary Material, Table A4). Fourth, we use Regression Adjustment (RA) and Inverse Probability

Table 3
Average Treatment effects from Propensity Score Matching using kernel matching.

Outcome	City	Treated	Controls	ATE	SE	t
SAT	Barcelona	69,673	46,228	-0.162***	0.009	-17.22
SAT	Madrid	28,739	19,952	-0.149***	0.015	-10.06
SAT	Milan	32,569	22,034	-0.221***	0.015	-14.97
SAT	Rome	63,677	40,580	-0.179***	0.011	-16.04
SAT	Lisbon	51,246	35,408	-0.155***	0.012	-12.55
SAT	Pooled Dataset	246,857	164,234	-0.174***	0.005	-33.68
SAT Rescaled	Barcelona	69,673	46,228	0.096***	0.011	9.08
SAT Rescaled	Madrid	28,739	19,952	0.117***	0.167	7.00
SAT Rescaled	Milan	32,569	22,034	0.075***	0.017	4.55
SAT Rescaled	Rome	63,677	40,580	0.084***	0.013	6.67
SAT Rescaled	Lisbon	51,246	35,408	0.099***	0.014	6.98
SAT Rescaled	Pooled Dataset	246,857	164,234	0.089***	0.006	15.37

*** $p < 0.01$. ** $p < 0.05$. * $p < 0.1$

Weighting (IPW) methods instead of PSM. These are alternative procedures to compute the effect of the change using the counterfactual scores that we would have had in the absence of the review system change that rebalance the data to account for differences in observables between consumers before and after (Horvitz and Thompson, 1952; Hirano et al., 2003). The results remain consistent using these alternative procedures (Supplementary Material, Table A5). Fifth, we repeated the analysis using an expanded sample that considers all the reviews in the period October 2018-January 2020. The OLS and PSM estimates for this larger sample are presented in Tables A6 and A7 in Supplementary Material and are very similar to the ones reported in Tables 2 and 3.

Our analysis lies on the assumption that, in the absence of the review system change, the mean overall scores would have remained unchanged. Despite we leave away seasonal differences by considering the same period before and after, the estimates could be biased if unobserved factors are responsible for the change in scores between the two periods. To explore this, we have performed a placebo exercise by generating two different fake treatments. The details and estimation results are shown in Tables A8 and A9 in Supplementary Material. Except for Barcelona, no significant differences in mean scores are detected before and after for the rest of cities, offering further evidence that the change in scores presented in Tables 2 and 3 can be attributed to the review system change.

6. Discussion

This work constitutes a novel endeavour in this research field, being among the first studies that conducts an econometric analysis to assess the impact of Booking.com review system change on hotel scores. Firstly, we offer a theoretical decomposition of the sources of the observed gap in scores before and after the platform replaced its multi-dimensional scoring system to a single-dimensional one for deriving the overall rating. Next, we exploit information for a total of 429,304 reviews for 1391 hotels in five of the most iconic European cities to draw causal inference about its effect on both *apparent* and *actual* scores. Based on both OLS regressions and Propensity Score Matching, we show that scores decreased on average by 0.15 points under the new system. However, once the scaling effect is adjusted, we document that the real scores increased by around + 0.09 points. These effects are quite similar across cities and remain consistent under different robustness checks.

Mellinas and Martin-Fuentes (2021) note that the change in Booking.com review system simply produces a scale change in scores since the underlying consumer satisfaction is likely to have remained unchanged. If this were the case, actual hotel scores after the scale adjustment should have been the same as in the previous system, at least on average. However, we provide evidence this is not the case.

Consistent with the framework developed in Section 3, the documented positive shift in *actual* scores is the sum of what we labelled the *weighting*, *variability,* and *omitted satisfaction* components. Allowing consumers to freely weight hotel attributes, to consider aspects other than the pre-determined six items, and to introduce their rates using a wider scale has led to positive increases in the real underlying satisfaction with hotel services revealed through consumers' scores. This suggests that either (i) attributes and service quality dimensions other than the six pre-determined items (e.g., aesthetics, sleep quality) seem to be positively valued by consumers, on average; (ii) consumers weight more positive than negative aspects when asked to provide an overall assessment of the satisfaction with their hotel stay; or a combination of both. In this respect, Nicolau et al. (2020b) show that customers' satisfaction is lower when the value for money is considered. The removal of value for money from the computation of hotel scores under the new system might explain the increased in actual scores, since hotel guests might weight other aspects more when making their evaluation. For instance, good language skills by the hotel staff could be more valued under the new valuation system (Nicolau et al., 2022).

From a broader perspective, our findings could be reconciled with the theory of unpacking effects (Van Boven and Epley, 2003) and the theoretical results presented in Schneider et al. (2021): asking people to rate their satisfaction with different domains before making an overall valuation make people less likely to express mild judgements and makes it easier for them to recall features that they would have not considered otherwise. In the Booking.com context, since satisfaction with the six items is still posited before the overall valuation, reminding people some hotel features but not others makes some of them more salient to memory and produces an upward shift in rescaled scores (i.e., *dimensional rating bias* as conceptualized in Schneider et al. (2021)). From this perspective, our results closely connect to those presented in Angelini et al. (2017), who show that overall life satisfaction scores are higher the question is posited after reporting satisfaction with different life domains.

Nonetheless, it is important to highlight that, in the short run, the change in the review system is expected to produce important negative effects for the sector. On the one hand, consumers infer unobserved quality based on observed scores left by previous consumers (Fernández-Barcala et al., 2010; Martín-Fuentes, 2016), which have suffered a decrease. Previous higher rates act as an anchor (Cicognani et al., 2021) that likely makes people judge the decline in hotel rates as a quality worsening instead of a mere scaling effect. Although one might consider this drop is common to all hotels, this is not the case. The hotel total overall score is computed as a weighted average of all scores received in the last three years. Depending on the percentage share of added reviews to the overall hotel score under the new system, some hotels (especially the most demanded) suffer a relatively larger decrease in observed rates than others. As a result, the pervasive effects of the review system change will not disappear until September 2022, when the hotel overall score will be computed using only the reviews under the new system (Amblee and Ullah, 2022; Mellinas et al., 2021).

7. Conclusions

7.1. Contributions

Our findings offer relevant theoretical and methodological contributions to the tourism literature. From a theoretical viewpoint, the paper expands existing work on online reputation in the hospitality industry by showing how the design and characteristics of crowd evaluation systems can produce non-neutral effects on both apparent and real hotel scores. Our findings fall within the choice architecture framework that postulates that the way choice options are presented to consumers affects their behavioural outcomes (Thaler, 2018; Schneider et al., 2021). Our conceptual framework is also a novel contribution since it offers a decomposition of the observed gap in Booking.com hotel scores, before and after the evaluation system change. Such gap is defined as the sum of what we call the *weighting*, the *scaling*, the *variability,* and the *omitted satisfaction* components.

From a methodological perspective, and consistent with a large literature on how to define measurement scales in tourism research (Dolnicar and Grün, 2013; Araña and León, 2013), our study emphasises the need for a careful design of satisfaction measurement in online platforms. As we show, moving from multi-dimensional to single-dimensional scoring systems has non-neutral effects on the reported level of satisfaction. Based on Propensity Score Marching on guests with similar profiles that stayed at the same hotel before and after the review policy change, our empirical analysis has offered causal evidence on how the review system change introduced by Booking.com has shifted apparent and actual scores at the individual review level. In this vein, our econometric strategy could be a promising methodology for future works interested in examining consumers' behavioural patterns following related external changes in online platforms' design.

7.2. Managerial implications

From a managerial perspective, the observed drop in scores is a matter of concern for hotel managers due to its implications on booking patterns through herding-behaviour (Figini et al., 2020). To avoid prospective guests to misjudge observed rates, awareness messages to inform people about the change in the review system could have been very informative. Beyond that, hotel managers could benefit from the use of web-scrapping procedures as the ones developed in this work to examine whether observed drops in scores are simply the result of the scale adjustment or reflect low satisfaction with previously neglected service quality dimensions. Although our analysis has been performed at the city level, our methods could be easily extended to the hotel level.

7.3. Limitations and avenues for future research

The paper has some limitations. First, our analysis lies on the assumption that mean scores would have remained unchanged if the review system had not been modified. Although we inspected this using a placebo exercise, we cannot completely rule out the possibility that external factors might have contributed to the detected difference in scores. This limitation is similar to the parallel trend assumption in difference-in-differences settings. Second, our data do not allow us to disentangle which part of the documented increase in actual scores is due to what we label as the weighting, the variability and omitted satisfaction components. Future studies should expand our analysis by trying to examine which of them dominates. Third, although we control for several relevant traveller's characteristics, our empirical analysis uses a full set of hotel fixed effects and does not distinguish between hotel size and category. Future research could expand our work by inspecting potential heterogeneity in the impact of the new review system based on hotel typology. Fourth, the worldwide outbreak of COVID-19 in February 2020 forced us to consider a relatively short time span that mainly used reviews left during the low season. Since ex-ante expectations and the quality of hotel services might be seasonal dependent, future work should investigate potential satisfaction heterogeneity between the high and the low season. As a final remark, given the high importance of hotel reviews on online platforms as quality cues, applied researchers using Booking.com are recommended to take into consideration whether their data are a mixture of reviews from the two valuation systems and, if so, make the corresponding scale adjustments.

Declaration of Competing Interest

None.

Acknowledgements

The authors wish to thank participants at the 8th Conference of the International Association of Tourism Economics for helpful comments and suggestions.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ijhm.2023.103493.

References

Akinci, S., Aksoy, S., 2019. The impact of service recovery evaluation on word-of-mouth intention: A moderated mediation model of overall satisfaction, household income and gender. Tour. Manag. Perspect. 31, 184–194.
Amblee, N., Ullah, R., 2022. Technique to harmonize Booking.com's dual rating systems. Ann. Tour. Res. 97, 103489.
Anagnostopoulou, S.C., Buhalis, D., Kountouri, I.L., Manousakis, E.G., Tsekrekos, A.E., 2020. The impact of online reputation on hotel profitability. Int. J. Contemp. Hosp. Manag. 32 (1), 20–39.
Angelini, V., Bertoni, M., Corazzini, L., 2017. Unpacking the determinants of life satisfaction: a survey experiment. J. R. Stat. Soc. A 180 (1), 225–246.
Araña, J.E., León, C.J., 2013. Correction for scale perception bias in tourist satisfaction surveys. J. Travel Res. 52 (6), 772–788.
Banerjee, S., Chua, A.Y.K., 2016. In search of patterns among travellers' hotel ratings in TripAdvisor. Tour. Manag. 53, 125–131.
Bateman, I., Munro, A., Rhodes, B., Starmer, C., Sugden, R., 1997. Does part-whole bias exist? An experimental investigation. Econ. J. 107 (441), 322–332.
Bigné, E., William, E., Soria-Olivas, E., 2020. Similarity and consistency in hotel online ratings across platforms. J. Travel Res. 59 (4), 742–758.
Borges-Tiago, M.T., Arruda, C., Tiago, F., Rita, P., 2021. Differences between TripAdvisor and Booking.com in branding co-creation. J. Bus. Res. 123, 380–388.
Boto-García, D., Escalonilla, M., Zapico, E., Baños-Pino, J.F., 2021. Scale heterogeneity in hotel guests' satisfaction relative to room rates. Appl. Econ. Anal. 29 (87), 208–225.
Cai, H., Chen, Y., Fang, H., 2009. Observational learning: evidence from a randomized natural field experiment. Am. Econ. Rev. 99 (3), 864–882.
Casaló, L.V., Flavián, C., Guinalíu, M., Ekinci, Y., 2015. Do online hotel rating schemes influence booking behaviors? Int. J. Hosp. Manag. 49, 28–36.
Cerulli, G., 2015. Econometric evaluation of socio-economic programs. Theory and applications. Springer.
Chen, C.W., 2017. Five-star or thumbs-up? The influence of rating system types on users' perceptions of information quality, cognitive effort and continuance intention. Internet Res. 27 (3), 478–494.
Chen, N., Li, A., Talluri, K., 2021. Reviews and self-selection bias with operational implications. Manag. Sci. 67 (12), 7472–7492.
Chen, P.Y., Hong, Y., Liu, Y., 2018. The value of multidimensional rating systems: Evidence from a natural experiment and randomized experiments. Manag. Sci. 64 (10), 4629–4647.
Choi, T.Y., Chu, R., 2001. Determinants of hotel guests' satisfaction and repeat patronage in the Hong Kong hotel industry. Int. J. Hosp. Manag. 20 (3), 277–297.
Cicognani, S. , Figini, P., Magnani, M. , 2021. Social influence bias in ratings: A field experiment in the hospitality sector. Tourism Economics, forthcoming.
Conti, G., Pudney, S., 2011. Survey design and the analysis of satisfaction. Rev. Econ. Stat. 93 (3), 1087–1093.
Cronqvist, H., Thaler, R.H., Yu, F., 2018. When nudges are forever. Am. Econ. Rev.: Pap. Proc. 108, 153–158.
Dolnicar, S., Grün, B., 2013. Validity measuring destination image in survey studies. J. Travel Res. 52 (1), 3–14.
Fernández-Barcala, M., González-Diaz, M., Prieto-Rodríguez, J., 2010. Hotel quality appraisal on the Internet: a market for lemons? Tour. Econ. 16 (2), 345–360.
Figini, P., Vici, L., Viglia, G., 2020. A comparison of hotel ratings between verified and non-verified online review platforms. Int. J. Cult. Tour. Hosp. Res. 14 (2), 157–171.
Gao, B., Li, X., Liu, S., Fang, D., 2018. How power distance affects online hotel ratings: the positive moderating roles of hotel chain and reviewers' travel experience. Tour. Manag. 65, 176–186.
Heckman, J.J., Ichimura, H., Todd, P., 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Rev. Econ. Stud. 64 (4), 605–654.
Heckman, J.J., Ichimura, H., Todd, P., 1998. Matching as an econometric evaluation estimator. Rev. Econ. Stud. 65 (2), 261–294.
Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71 (4), 1161–1189.
Hörner, J., 2002. Reputation and competition. Am. Econ. Rev. 92 (3), 644–663.
Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. 47 (260), 663–685.
Jann, B., 2017. KMATCH: Stata module for multivariate-distance and propensity-score matching. Statistical Software Components, Boston College Department of Economics.
Kim, J.M., Han, J., 2022. Impact of the length of stay at hotels on online reviews. Int. J. Contemp. Hosp. Manag. 34 (4), 1249–1269.
Kim, J.M., Liu, J., Yousaf, S., 2022. Does change in the scoring system impact service evaluation? Evidence from Booking.com. Int. J. Contemp. Hosp. Manag. DOI 10.1108/IJCHM-01-2022-0075.
Kim, M., Lee, S.M., Choi, S., Kim, S.S., 2021. Impact of visual information on online consumer review behavior: evidence from a hotel booking website. J. Retail. Consum. Serv. 60, 102494.
Kim, M.G., Mattila, A.S., 2010. The impact of mood states and surprise cues on satisfaction. Int. J. Hosp. Manag. 29 (3), 432–436.
Lee, Y.J., Hosanagar, K., Tan, Y., 2015. Do I follow my friends or the crowd? Information cascades in online movie ratings. Manag. Sci. 61 (9), 2241–2258.
Leung, D., Law, R., van Hoof, H., Buhalis, D., 2013. Social media in tourism and hospitality: a literature review. J. Travel Tour. Mark. 30, 3–22.
Luca, M., Zervas, G., 2016. Fake it till you make it: reputation, competition, and Yelp review fraud. Manag. Sci. 62 (12), 3393–3672.
Lyu, J., Khan, A., Bibi, S., Chan, J.H., Qi, X., 2022. Big data in action: an overview of big data studies in tourism and hospitality literature. J. Hosp. Tour. Manag. 51, 346–360.
Magnani, M., 2020. The economic and behavioral consequences of online user reviews. J. Econ. Surv. 34 (3), 263–292.
Mariani, M., 2020. Big data and analytics in tourism and hospitality: a perspective article. Tour. Rev. 75 (1), 299–303.
Mariani, M., Predvoditeleva, M., 2019. How do online reviewers' cultural traits and perceived experience influence hotel online ratings? An empirical analysis of the Muscovite hotel sector. Int. J. Contemp. Hosp. Manag. 31 (12), 4543–4573.
Mariani, M., Borghi, M., Gretzel, U., 2019. Online reviews: differences by submission device. Tour. Manag. 70, 295–298.

Mariani, M.M., Borghi, M., 2018. Effects of the Booking.com rating system: bringing hotel class into the picture. Tour. Manag. 66, 47–52.
Martin-Fuentes, E., Mellinas, J.P., Parra-Lopez, E., 2020. Online travel review rating scales and effects on hotel scoring and competitiveness. Tour. Rev. 76 (3), 654–668.
Martín-Fuentes, E., 2016. Are guests of the same opinion as the hotel star-rate classification system? J. Hosp. Tour. Manag. 29, 126–134.
Martín-Fuentes, E., Mellinas, J.P., 2018. Hotels that most rely on Booking.com - online travel agencues (OTAs) and hotel distribution channels. Tour. Rev. 73 (4), 465–479.
Martín-Fuentes, E., Mateu, C., Fernandez, C., 2018. Does verifying uses influence rankings? Analyzing Booking.com and TripAdvisor. Tour. Analysis 23 (1), 1–15.
Mayzlin, D., Dover, Y., Chevalier, J., 2014. Promotional reviews: an empirical investigation of online manipulation. Am. Econ. Rev. 104 (8), 2421–2455.
Mellinas, J.P., Martin-Fuentes, E., 2021. Effects of Booking.com's new scoring system. Tour. Manag. 85, 104280.
Mellinas, J.P., Nicolau, J.L., 2020. Let's hook up fast! Hotel reviews and Wi-Fi flaws. Ann. Tour. Res. 80, 102842.
Mellinas, J.P., Soledad-María Martínez, M.D., Bernal-García, J.J., 2015. Booking.com: the unexpected scoring system. Tour. Manag. 49, 72–74.
Mellinas, J.P., Nicolau, J.L., Park, S., 2019. Inconsistent behavior in online consumer reviews: the effects of hotel attribute ratings on location. Tour. Manag. 71, 421–427.
Moreno-Perdigón, M.C., Guzmán-Pérez, B., Ravelo-Mesa, T., 2021. Guest satisfaction in independent and affiliated to chain hotels. Int. J. Hosp. Manag. 94, 102812.
Naumzik, C., Feuerriegel, S., Weinmann, M., 2022. I will survive: predicting business failures from customer ratings. Mark. Sci. 41 (1), 188–207.
Nicolau, J.L., Mellinas, J.P., Martín-Fuentes, E., 2020a. The halo effect: a longitudinal approach. Ann. Tour. Res. 83, 102938.
Nicolau, J.L., Mellinas, J.P., Martín-Fuentes, E., 2020b. Satisfaction measures with monetary and non-monetary components: Hotel's overall scores. Int. J. Hosp. Manag. 87, 102497.
Nicolau, J.L., de Carlos Villamarín, P., Alén, E., Pérez-González, A., 2022. Asymmetric effects of extreme-moderate online reviews in the language-satisfaction relationship. Tour. Manag. 91, 104524.
Nieto-Garcia, M., Resce, G., Ishizaka, A., Occhiocupo, N., Viglia, G., 2019. The dimensions of hotel customer ratings that boost RevPAR. Int. J. Hosp. Manag. 77, 583–592.
Phillips, P., Antonio, N., de Almeida, A., Nunes, L., 2020. The influence of geographic and psychic distance on online hotel ratings. J. Travel Res. 59 (4), 722–741.
Pizam, A., Shapoval, V., Ellis, T., 2016. Customer satisfaction and its measurement in hospitality enterprises: a revisit and update. Int. J. Contemp. Hosp. Manag. 28 (1), 2–35.
Preston, C.C., Colman, A.M., 2000. Optimal number of response categories in rating scales: realiability, validity, discriminating power and respondent preferences. Acta Psychol. 104, 1–15.
Rabin, M., Schrag, J.L., 1999. First impressions matter: a model of confirmatory bias. Q. J. Econ. 114 (1), 7–82.
Racherla, P., Connolly, D.J., Christodoulidou, N., 2013. What determines consumers' ratings of service providers? An exploratory study of online traveler reviews. J. Hosp. Mark. Manag. 22 (2), 135–161.
Radojevic, T., Stanisic, N., Stanic, N., 2015. Solo travellers assign higher ratings than families: Examining customer satisfaction by demographic group. Tour. Manag. Perspect. 16, 247–258.
Rita, P., Ramos, R., Borges-Tiago, M.T., Rodrigues, D., 2022. Impact of the rating system on sentiment and tone of voice: A Booking.com and TripAdvisor comparison study. Int. J. Hosp. Manag. 104, 103245.
Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Rottenstreich, Y., Tversky, A., 1997. Unpacking, repacking, and anchoring: advances in support theory. Psychol. Rev. 104 (2), 406–415.
Rubin, 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66, 688–701.
Schneider, C., Weinmann, M., Mohr, P.N.C., vom Brocke, J., 2021. When the stars shine too bright: the influence of multidimensional ratings on online consumer ratings. Manag. Sci. 67 (6), 3871–3898.
Serra Cantallops, A., Salvi, F., 2014. New consumer behavior: a review of research on eWOM and hotels. Int. J. Hosp. Manag. 36, 41–51.
Shen, X., Pan, B., Hu, T., Chen, K., Qiao, L., Zhu, J., 2021. Beyond self-selection: the multilayered online review biases at the intersection of users, platforms and culture. J. Hosp. Tour. Insights 4 (1), 77–97.
Smironva, E., Kiatkawsin, K., Lee, S.K., Kim, J., Lee, C.H., 2020. Self-selection and non-response biases in customers' hotel ratings - a comparison of online and offline ratings. Curr. Issues Tour. 23 (10), 1191–1204.
Sparks, B.A., Browning, V., 2011. The impact of online reviews on hotel booking intentions and perception of trust. Tour. Manag. 32, 1310–1323.
Štumpf, P., Vojtko, V., McGrath, R., Rašovská, I., Ryglová, J., Šácha, J., 2022. Destination satisfaction comparison excluding the weather effect. Curr. Issues Tour. 25 (15), 2404–2421.
Taecharungroj, V., Mathayomchan, B., 2019. Analysing TripAdvisor reviews of tourist attractions in Phuket, Thailand. Tour. Manag. 75, 550–568.
Thaler, R.H., 2018. From cashews to nudges. Am. Econ. Rev. 108 (6), 1265–1287.
Thaler, R.H., Benartzi, S., 2004. Save more tomorrow: using behavioral economics to increase employee saving. J. Political Econ. 112 (S2), S164–S187.
Van Boven, L., Epley, N., 2003. The unpacking effect in evaluative judgements: when the whole is less than the sum of its parts. J. Exp. Soc. Psychol. 39 (3), 263–269.
Vermeulen, I.E., Seegers, D., 2009. Tried and tested: the impact of online hotel reviews on consumer consideration. Tour. Manag. 30, 123–127.
Viglia, G., Minazzi, R., Buhalis, D., 2016. The influence of e-word-of-mouth on hotel occupancy rate. Int. J. Contemp. Hosp. Manag. 28 (9), 2035–2051.
Zhang, C., Xu, Z., Gou, X., Chen, S., 2021. An online reviews-driven method for the prioritization of improvements in hotel services. Tour. Manag. 87, 104382.