



STEFAN THOMKE

DANIELA BEYERSDORFER

Booking.com

Our company has thought a lot about how to make the booking of accommodations informative and intuitive for customers. We never stop improving user experiences. Everything is a test.

— Gillian Tans, CEO, Booking.com

Gillian Tans, CEO of Booking.com, and David Vismans, Chief Product Officer, scratched their heads over an experiment that one of the company's managers was going to run with millions of customers. The test involved a new customer interface that had little resemblance with the company's hugely popular landing page, which had propelled it to the largest accommodation-booking platform in the world (see **Exhibit 1**). It was right before Christmas and one of the company's busiest travel periods. The experiment's web interface consisted of a blue background, a simple Google-like search box in the center, and booking options that included accommodations, flights and rental cars. Neither Tans nor Vismans believed that this experiment would improve customer conversion—website visitors that would make a booking—which was the company's most important performance metric. If anything, it could lead to massive confusion and defections among its loyal customer base, as they would not recognize the interface. As much as Tans and Vismans were proud of the company's "test everything" ethos, which empowered employees to launch experiments without management approval, they now wondered if this experiment would go too far.

Booking.com (hereafter Booking) had grown from a small Dutch startup to one of the world's largest online travel companies. Based in Amsterdam, its headquarters was spread over 10 buildings to accommodate employees from more than 100 nationalities. Its team-oriented culture emphasized autonomy and empowerment; new recruits were selected for their experimentation mindset, which included innovative thinking, fast-decision making, fearlessness, and a willingness to openly share failures. Booking prided itself on connecting travelers with the world's largest selection of hotels and places to stay. Every day, more than 1.5 million room nights were booked on its platform and it offered more than 1.6 million properties in 227 countries. To fulfill its mission to "empower people to experience the world" it invested heavily in digital technology to "take the friction out of travel." Booking was known for its relentless focus on customer-centric product development via online experiments, notably A/B testing, and for the way it had democratized experimentation throughout its organization. On any day, its staff ran more than 1,000 rigorous tests on its website, servers, and apps to optimize customer experiences. With quadrillions (millions of billions) of landing page permutations running live, customers booking a room on its website were all part of Booking's experimentation ecosystem.

Professor Stefan Thomke and Associate Director Daniela Beyersdorfer (Europe Research Center) prepared this case. It was reviewed and approved before publication by a company designate. Funding for the development of this case was provided by Harvard Business School and not by the company. HBS cases are developed solely as the basis for class discussion. Cases are not intended to serve as endorsements, sources of primary data, or illustrations of effective or ineffective management.

Copyright © 2018 President and Fellows of Harvard College. To order copies or request permission to reproduce materials, call 1-800-545-7685, write Harvard Business School Publishing, Boston, MA 02163, or go to www.hbsp.harvard.edu. This publication may not be digitized, photocopied, or otherwise reproduced, posted, or transmitted, without the permission of Harvard Business School.

Online Travel Industry

The online travel industry comprised primarily travel e-commerce and review sites. Travel e-commerce sites allowed customers to purchase travel products such as hotels, flights, and rental cars, either directly through a travel company's website (e.g., Lufthansa airline) or through an online travel agency (OTA) acting as an intermediary.¹ OTAs had agreements with hotels and other suppliers of travel products to purchase some of their inventory and then allowed customers to book those products on their website or through mobile apps. Travel review websites, such as TripAdvisor, allowed customers to share their experience with travel products, for example by rating a hotel stay, and often generated revenue via advertisements on their website. Travelers worldwide were increasingly relying on travel review sites when booking.²

In 2017 global online travel sales generated US\$630 billion (up 11.5% from 2016) and were expected to reach \$818 billion by 2020.³ Expedia Inc., The Priceline Group (Booking's owner),⁴ and China's Ctrip had become the largest travel agents worldwide in bookings and sales.⁵ TripAdvisor ranked first in number of users.⁶ The four companies had driven OTA consolidation to expand market share and were now competing against direct suppliers such as hotels.⁷ The OTAs themselves were challenged by new entrants, such as the peer-to-peer site AirBnB, and the search-engine giant Google.⁸ Google had launched a Hotel Finder tool in 2011, which by 2016 had grown into a full-blown hotel search service. It had also added flight search with links to airline websites, allowing travelers to compare and book flights and hotels without having to go through OTAs. OTAs, which relied heavily on Google for customer traffic, fought back by increasing advertising spending. Priceline and Expedia spent over \$6 million in 2016,⁹ and increased advertising spending in 2017. By 2017, Google was expected to generate \$14 billion in revenue from its travel business. Analysts speculated that Amazon could be among the companies that will enter the travel sector.¹⁰

Building Scale

In November 1996 recent university graduate Geert-Jan Bruinsma founded Bookings.nl in Amsterdam. Bruinsma had become fascinated with the nascent Internet and its opportunities to start new businesses. By 1997, the website went live with 10 hotels, allowing customers to reserve rooms online. Wanting to become number one for Amsterdam hotels, Bruinsma struck a deal with the two leading Amsterdam hotel sites (Channels.nl and Amsterdam Hotel Guide), which put his hotels behind their sites. Initially, many customers came from the U.S. as Internet access was still scarce among Europeans. A first employee joined in 1998.¹¹ Vismans explained how Booking benefited from Google's growth, "The Internet created a lot of opportunities. Startups were like surfers in the water waiting for the right wave, and those with the best execution managed to ride it. Among the opportunities was Google's launch of AdWords in 2000, a real game changer as it revealed customer intent. Whoever searches for 'hotel Amsterdam' is clearly a potential customer and not someone trying to build a hotel here."

Bruinsma continued to work on the landing page and learned about search engine optimization (SEO). And as Google grew bigger, Booking scaled its business along with it. Tans recalled:

When I joined Booking in 2002, my family thought I was crazy. The company was still tiny; I was the 7th employee. We had to fix so many things. Many companies start with a nice product and market it all over the world. Booking did the opposite. We had a basic product and then worked hard to get it right for customers. But figuring out what they like is hard. We got it wrong so many times. For example we thought they'd like to see videos of hotels and then realized they did not look at them. Or we believed customers

when they said they'd go only for price and then saw them act differently. Starting 2004, we ran simple tests to learn which options they prefer, initially just a few times per day without any big technology behind it, and then built the product based on their preferences. We grew like this, without any marketing or PR, just testing what our customers liked.

Vismans added, "I believe Booking was among the first in the travel industry to become test and data-driven. It's because we learned that intuition is wrong most of the time, especially online where you have no experience with customer behavior and matching supply and demand."

Tans thought that the Dutch origins had been beneficial, "We operated only in Holland when I started. Our country is so small but Dutch people travel abroad quite a lot. To follow demand we built an international platform, while our competitors in larger countries focused on their home markets." One of Booking's first expansion decisions was the selection of an office location in Germany. Vismans explained, "Conventional wisdom suggested to start in Berlin where you expect most Dutch tourists. But we decided to check which city comes up first in customer searches. It turned out to be a village called Winterberg, a ski paradise for the Dutch. So we followed the data and open our first office there." To expand efficiently, Booking focused on building a universal yet simple product. Vismans summed up the company's success factors: "Growing through key enablers, building your product through experimentation, and following demand, together with the insights this generates for management, made the company realize that they hit on something big; that if they just executed really well, they were going to be in a really good place."

Booking operated with an "agency model," in which customers booked rooms on its website and paid directly to the hotel. Tans noted, "With this model you can scale very fast, you don't need a payment infrastructure and the hotels manage the inventory. And it's what European customers prefer. They are not used to paying up-front and want flexibility." Booking's primary revenues came from commission fees (averaging 15% globally) for non-cancelled rooms, collected once per month by sending a list of its reservations to respective hotels. In the early 2000s, competitors such as US-based Expedia (launched 1996) entered the European market but struggled. The new entrants operated with a "merchant model," in which they bought room contingents from hotels and collected payments at the time of a reservation, making it harder for customers to cancel. Tans said, "Our competitors were more like travel agents, with flights and other options for which that merchant model makes more sense. And their margins and cash flow benefit from the earlier collection of money."

By 2005 Booking was on its way to becoming the European market leader. Its success caught the attention of US-based The Priceline Group, which bought Booking for a mere US\$133 million in cash and gave management the budget and mandate to scale further.¹² Around the same time, Booking completed the development of an experimentation platform that allowed it to scale testing as well. Adrienne Enggist, Director of Product Messaging, recalled, "I came from small businesses where CEOs launched a big product redesign every six months, and by the time you rolled it out, it was hard to figure out what worked and what did not work. Here the team was small, fitted on one floor, and it was exciting to see everyone take risks, push small changes very quickly, and use experiments to measure the impact. The idea was the more lines in the water, the more fish you can catch. And while people were less savvy about experimentation than today, it was easy to get things done."

In the following years Booking grew reservations and revenues quickly and stayed focused on accommodations. Tans noted, "We always felt that accommodation is what's crucial for a trip. So you better get the booking process right, and for a long time, we felt that there was so much work to do. Many competitors diversified too early. But sometimes it's just as important what you don't do. We always felt that if we make the best product, and build the fastest execution machine, we are going to

win.” As Booking’s staff grew, the company also expanded its global hotel inventory. To differentiate its user experience, the company invested in a scalable “Content Agency” for languages, initially using translators and increasingly machine learning, to feature its contents in a growing number of languages. In 2014, the rapid growth of experiments triggered an overhaul of its testing platform and standardization of its methods.

To grow inventory on its platform, Booking had built a global network of hotels and accommodation providers, so-called partners. Enggist explained, “We are a two-sided platform. One of our interesting challenges is our position as a way for both parties to connect; for a guest to find the hospitality provider, and for our supplier partner to optimally display offers.” From the start, Booking had made it easy for new partners to join and display their rooms via its extranet, app, or data connection, rather than having to go through lengthy negotiations and wait for OTAs to put rooms online. Partners could connect to the platform and manage their inventory, uploading the number of rooms they wanted to make available, at the price set by them. To recruit and support partners, Booking had 200 offices worldwide, with 4,000 account managers serving as local ambassadors, and sales support for new partners. While the majority of new sign ups occurred through an automated web link, larger partners still valued the personal interaction. Booking was one of the several sales channels for its partners. The company’s added value consisted of offering hotels a popular platform on which they could market excess inventory worldwide. Booking also helped property owners run their business more effectively through analytics (information on demand, pricing, aggregated competitor statistics, guest reviews, etc.). Unlike TripAdvisor, Booking had a “closed” review system, in which only former guests of a property could leave a review. Good review scores helped properties to move up in default search rankings and scores of eight or above gave them the option to access a preferential partner program.

In 2017, in response to AirBnB and other entrants, Booking increased its offering of “alternative accommodations” to 1.2 million homes and apartments (up 53% over 2016).¹³ It also ran tests in multiple markets with “in-destination experiences,” such as tickets for tourist attractions.¹⁴ By December 2017, Booking’s offerings included over 1.6 million properties (hotels, apartments, vacation homes, bed & breakfasts, and more) in 120,000 destinations. Its website and mobile app were available in 43 languages. Booking employed 15,000 people in 199 offices in 70 countries. One third of them, and most corporate functions, were based in Amsterdam; the rest in a small tech center in Israel, a product & marketing center in Shanghai, and call centers all over the world.

The Priceline Group’s 2017 year-end results had shown significant growth across all sectors. Through its six brands—Booking.com, Priceline.com, Kayak, Agoda.com, Rentalcars.com, and OpenTable—it had generated revenue of \$12.7 billion (up 18% from 2016). Industry observers estimated that about 70% to 80% of that revenue was generated by Booking alone. The Priceline Group’s gross travel bookings had been \$81.2 billion (up 19%) and gross profit of \$12.4 billion (up 21%).¹⁵ In December 2017, The Priceline Group’s market cap was approaching to \$90 billion. Again, analysts attributed most of its financial success to Booking. (See **Exhibit 2** and **3** for key figures and financials.)

A/B Testing

The company’s focus on optimizing customer experiences had remained unchanged since its early days. Vismans explained, “If you want to be successful, you need to offer a great customer experience. This has to be your sole focus when developing products. Whenever they come in contact with your website, it needs to be more satisfying than with competitors, so they come back.” To figure out what customers found satisfying, its developers continuously tried out ideas to improve the product experience through controlled online experiments, augmented with qualitative research. Failure was

accepted as a normal byproduct, as long as it accelerated the improvement process. Senior Product Owner of Experimentation Lukas Vermeer noted, “We call this evidence-based, customer-centric product development. All our product decisions are based on reliable evidence about customer behavior and preferences. We believe that controlled experimentation is the most successful approach to building products that customers want.”

The simplest kind of a controlled experiment was an A/B test (see **Exhibit 4** for examples). In this test, the experimenter sets up two experiences: “A”, the control, is usually the current system and considered the “champion,” and “B”, the treatment, is a modification that attempts to improve something – the “challenger.” Customers are randomly assigned to the two experiences, and key metrics are computed and compared. Online, the modification could be a new feature, a change to the user interface (such as a new layout), a back-end change (such as an improvement to an algorithm), or a different business model (such as a discount offer). Whatever aspects of performance teams cared most about – be it sales, repeat usage, click-through-rates, conversion, or time users spend on a site – Booking could use A/B tests to learn how to optimize them.¹⁶ Vismans explained, “If we need to create a ‘book’ button, we want to understand what the color of the button should be. So we create two versions of the website, one with a yellow button and the other with a blue one, to test them live on millions of customers. We will use the color that attracts the most bookings. Our customers decide where to take the website, not our managers.”¹⁷

Deciding if a “challenger” was winning against the “champion” wasn’t always easy. Managers had to agree on key performance indicators (KPIs), or metrics, they would watch to judge performance. Booking’s primary metric was *user conversion*, measured as bpd (bookings per day). But with a growing business and maturing product, it was important to measure post-booking behavior as well. Tans noted, “The issue with bpd is that it is short term and does not pick up on problems that may arise later. Let’s say our cancellation policy gets less clear; customers pay without noticing and then complain to customer services afterwards. Those longer-term signals are harder to pick up in experiments but we try to take them into account, even if that means small hits on bpd.” While about 80% of its staff focused on conversion, teams were free to include other metrics in their experiments.

Booking had learned early on that it could not trust intuition and assumptions. “We see evidence every day that people are terrible at guessing. Our predictions of how customers will behave are wrong nine out of ten times,” Vermeer said. Intuition had proven unreliable in all areas, be it to guess which colored button users prefer or which functionalities they value. Tans recalled, “For example, we mistakenly believed that customers would like hotel offers packaged with other products since travel brochures are full of them. Or we thought that customers would want a chat line helping them through the booking process. Neither idea worked during our tests. It’s how you learn.” Vismans added, “We have done [it] this way for nine years and it’s very effective in building something that customers find most valuable or easy to use. We follow what the majority wants. And if you fail fast, you can try a lot of things.”¹⁸ Vermeer agreed: “It’s like some sort of rapid prototyping. As a digital company we have many touchpoints with customers to test and optimize.”

One source of inspiration for touchpoint tests was qualitative insights into customer behavior. To find them, Booking ran an in-house user experience (UX) lab with 45 researchers. They utilized feedback reports, online surveys, usability tests, street testing, and home visits to study how customers used Booking products in their daily routines. Consumer psychologist Gerben Langendijk explained, “Our product teams can order funnel tests in our lab, where they observe how people navigate through the website, what they think, and how they struggle. It’s very powerful for teams to see this, especially when they think a new function is obvious but users don’t understand it. Tests at users’ homes show us how they behave with our product in their own environment, spending their own money. We run tests on the street, in bars, and in cafés here in Amsterdam. We show mockups, so people can try a new

user interface. We also go abroad to focus on specific markets and capture cultural preferences. For our partners we look into how we can improve their supplier experience.” The resulting data was made available to teams, so they could brainstorm new features, improve existing ones, and solve user issues.¹⁹

Another source of insights was Booking’s customer service department, which was available 24 hours/7 days a week for assistance and support in 43 languages. Customers could sort out many issues online, like change or cancel a booking, but they could also call a live person. Booking’s customer service centers answered about 14 million calls each year and had noticed that customers’ expectations of product quality had steadily gone up. The service department forwarded relevant feedback to developers, so they could use it for new experiments.²⁰ Principal data scientist Onno Zoeter noted, “They provide important feedback about the back end of customer experiences, on how our product holds up in the longer term. We invest a lot in customer service; remote call centers have the same type of desks and chair as our CEO and get flown in for the annual Booking meeting and party in Amsterdam.”

Vismans viewed Booking’s competitive advantage as executing its business model through large-scale testing. “We buy demand from Google through advertising spending, convert that demand to bookings, add a positive return-on-investment (ROI), and then source supply based on that demand. And since we have a KPI that correlates with our bottom line, we ask everyone to experiment as much as they can. The only requirement is that all changes have to be tested. So you get the cumulative effect of many small changes that, over time, nobody can compete with anymore.” Vismans continued:

We have our own version of Amazon’s flywheel concept (see **Exhibit 5**). It’s a virtuous cycle with network effects, where each component is an accelerator. Invest in any one of them, and as the wheel spins, it benefits all and generates growth. For us it starts with a great customer experience. Through A/B testing we improve the product experience, which drives up conversion. The more people and conversion we get, the faster the wheel spins and the larger the marketing ROI and traffic, which leads to more partners wanting to be on our platform and more leverage for us. This, in turn, means a broader selection at cheaper prices and the best service, which again leads to greater customer experiences. It’s a “growth leads to growth” model. You can’t neglect any aspect of it; if conversion breaks down, you can’t fulfill the contract anymore, so you need to keep your eye on the metrics. Whatever you start, you need to define metrics and then A/B test against it. If you want partners to give you more availability, you start testing. In the end, your whole business model becomes testable. But you need to understand the strategy first. If you do A/B testing without understanding how your network effects are connected you’re just running around like a headless chicken.

The Experimentation Organization

I can come up with an idea over breakfast, bike into work, and have it implemented and live well before lunch. I’ve never worked anywhere else with so much freedom to validate my ideas.

— User Experience Copywriter, Booking.com

By 2017, Booking ran about 1,000 controlled experiments concurrently. They were launched and analyzed by employees from all departments, and ran across all products, from the website to mobile apps, on tools used by partners to customer service lines, and on internal systems. About 80% of tests ran on the “core” —everything linked to the actual accommodation booking experience—resulting in quadrillions of different landing page variants being live simultaneously. Customers were randomly

distributed between controls and variants, and most experiments were subjected to most customer traffic. Director of Design Stuart Frisby noted, “This makes for an astronomical number of permutations. It also means that two customers in the same location accessing Booking’s website are unlikely to see the same version.” Senior Product Director Andrea Carini added:

We have a philosophy of testing as much as we can live with customers, and some tests take several iterations or are revisited later, which adds to these high numbers. Everything is tested, from entire redesigns and infrastructure changes to small bug fixes. If I have a software bug, I want to make sure that my fix improves the user experience. So we split test the bug, keep it in the A group and put the fix in B to make sure that the new code actually solves the problem, and doesn’t negatively impact customer metrics.

Booking had built an in-house experimentation platform to ensure tests were easily doable by everyone but also rigorous in their execution (see **Exhibit 6**). The company had a dedicated “core experimentation team” of seven, led by Vermeer and part of the core infrastructure department, that took care of the experimentation infrastructure and tools, and provided training and support to the whole organization. Vermeer noted, “My team’s mission is to enable all of our employees to run experiments autonomously.” Five “satellite” support teams were placed directly in Booking’s product departments; other support teams moved to partners and customer service to help them ramp up experimentation as well. Vermeer explained, “Teams specialize in a product area, sit on the same floor, and go to the same meetings.” Other teams specialized in improving the experimentation platform or explored advanced statistical methodologies. The support teams divided their time between “helpdesk support” for experiments running in their departments, preparation of information for management on how experiments were doing, and improvements of tools and metrics. Vermeer emphasized the importance of autonomy: “If a team thinks they need reminder emails for their tests, they are free to build them. And if that feature works well and is requested by other teams we pull it over to the core and centralize it for everyone. Each team reports to its department but I rotate daily to see them. We also have regular meetings between them and quarterly one-day off-site events where we exchange best practices.”

Booking’s platform was designed to make experimentation accessible to everyone. To encourage openness, it offered a central searchable repository of past experiments, with full descriptions of successes, failures, iterations, and final decision. Standard templates allowed the setup of experiments across departments and products with minimal ad-hoc work, and processes like user recruitment, randomization, recording of visitor’s behavior, and reporting were automated behind a set of application programming interfaces (APIs). To make experiments trustworthy, the validity of data was monitored by computing a set of common metrics in two entirely separate data pipelines, maintained by engineers to quickly detect bugs. Several safeguards were built into the platform, allowing experiments to be monitored by both owners and the community, before and during their execution. Vermeer explained, “Somewhat ironically the centralizing of our experimentation infrastructure is what makes our organizational decentralization possible. Everyone uses the same tools. This fosters trust in each other’s data and enables discussion and accountability. While some companies like Microsoft, Facebook, or Google may be more technically advanced in areas like machine learning, our use of simple A/B tests makes us more successful in getting all people involved; we have democratized testing throughout the organization.” Frisby added, “About 75% of our 1,800 technology and product people actively use the experimentation platform, which is huge, and now we’re including partners and customer services as well.”

“The people who thrive here are curious, open minded, eager to learn and figure things out, and are okay with being proven wrong,” Vermeer emphasized, “Some join because they want to work on a website with a lot of traffic, where they can validate their ideas with data.” Vermeer’s group provided

the training for new recruits, “People expect to learn about the tool, but for the first few hours we talk to them about the scientific method, and then about experiments, hypotheses, statistical terminology, design of experiments, ethics, compliance, and so on.” Newcomers were paired with a senior staff member who explained the work in more detail, introduced the platform, and analyzed experiments and related decisions. New recruits also had access to all tools and could gain hands-on experience early on. A developer noted, “Experimentation at Booking is a constant evolution. I sometimes laugh at the experiments I did four years ago for the lack of secondary metrics, and to this day, we are still pushing the bar higher, innovating on how we conduct experiments.”²¹

Governance and Culture

Booking was organized in four main departments: products (the largest), followed by partner services, customer service, and core infrastructure (see **Exhibit 7**). The company’s structure had remained relatively flat, with just a few senior VPs, product owners and technology managers, and with decisions pushed down as far as possible. Carini noted, “Not everything is neatly organized and not everyone has clear reporting lines. These are the typical stretch marks of a company growing at an exponential rate. Booking is 21 years old but most employees joined in the last eight years. It’s also not efficient to have a neat structure. How can you innovate and react in our fast-moving industry if you sit in a neatly boxed place and wait to be told what to do?” Vermeer added, “Some people struggle with the flat structure because there can be little room to move up. However, anybody can do anything. Teams and individuals have a lot of responsibility and people move around, which keeps it interesting and allows them to see different parts of the customer journey.” Booking did quarterly performance reviews for all employees, which included feedback from managers and peers, and a self-assessment.

Throughout the company, employees were organized in multidisciplinary teams of six to eight people. Each team had a product owner (e.g., billing, landing pages) who was responsible for the product roadmap from a business perspective. The rest of the team was made up of tech people—engineers and designers—tasked with coding and implementing ideas. This often included a front-end and back-end developer, designer, copywriter, researcher, and data analyst. Anybody on a team could launch an experiment; however, 90% of tests came from teams rather than individuals. Carini noted, “Usually teams work together to launch a test. The product owner comes up with the problem, the engineers decide the variables, and then all work together on the right hypothesis, execution, and iteration. Everyone is familiar with testing so you can have good conversations.” Typically designers spent about 75% of their time on designing experiments and 25% on research and professional development. Senior employees spent a lot of their time coaching. Frisby added, “I develop tools like reusable lists, so other designers don’t have to build them from scratch. Since most experiments fail, we want them to be designed and executed with the least effort and time but also with the best quality. Proven, stress-tested tools can help with that.”

Teams were encouraged to run as many experiments as possible. Frisby continued, “Anyone can do anything, play with whatever they want. Nothing is sacred, except for legal constraints, fair display of properties, and those kinds of things.” Vismans noted, “Once you have decided that testing is the right way for your organization to build products and have the right metrics, you have no choice than to give everyone autonomy. It’s the only efficient way to unlock team creativity. The success rate of experiments is so low that you need to try a lot. Senior management directives that interfere with innovation would only slow the process. It’s close to anarchy. Or better, it is organized chaos. The KPIs and objectives ensure that people know what and how to test.” Carini clarified, “Obviously we also have our shared company values, a formula for how we do things, so we know people would not do completely crazy things, like putting illegal content live. Values are: be data driven in your decisions, always put the customer first, etc.” (See **Exhibit 8**.)

New recruits were granted autonomy very quickly. Senior Product Owner Willem Isbrucker recalled, “When I joined, I was baffled by the level of trust. I could make decisions on experiments from day one and take full control of follow-ups within a week. Let’s say you want to make the website pink. If you have any evidence showing that this may be good for users, then you can test. That’s a huge difference to my previous employers. When I realized that I could run daily tests on millions of people, I was extremely happy.”

The high level of autonomy also came with challenges. One risk was that teams and individuals could break something on Booking’s high traffic website, which could result in a crash. Moreover, in such a decentralized bottom-up organization, each team had to set its own direction and figure out what users problems they wanted to solve. For employees, this meant huge responsibility. Isbrucker continued, “There is nowhere to hide here, no scapegoat you can put the blame on if you don’t find user problems and how to solve them or if you break something.” Debates were encouraged and people reached out to colleagues if they saw anything they found questionable or did not agree with. Anybody could stop any experiment at Booking, though, as Vermeer noted, “In reality, it happens rarely. Usually you would approach a team if you saw a problem, for instance, by asking them if they noticed that they were bleeding 2% conversion and if they are on top of it. Pushing the stop on someone’s test was seen as very aggressive, the nuclear approach. It’s only done if there is no other option; say you are alone at the office at night and there is an incident in some part of the world that requires an immediate stop.”

One issue that had fueled vigorous debates was the use of persuasion techniques. For example, product pages featured messages, such as “please book now or you will lose this reservation,” or “in high demand,” or “only three rooms left.” While these messages were originally intended to inform consumers about availability, some people perceived the messages as conveying scarcity and urgency. Critics argued that such messages could mislead customers to think there were only three rooms left in the entire hotel, when in fact the three rooms were about a hotel’s allocation to Booking. After regulators got involved, Booking rectified the message to “only three rooms left on our platform.” Ethics debates regularly flared around whether the increasing use of such techniques was in the best interest of customers.

Experiments showed that this kind of messaging worked – the conversion metric improved – so customers did respond positively. Leveraging psychological techniques was also an easy way for new employees to show a quick test win. Psychologist Langendijk explained, “When teams ask me to work on persuasive elements, I explain first that the best persuasion is having a great product. We need to see where such elements make sense, for example, when an experienced visitor identified the right hotel and is about to book, and where they may hurt people, particularly first time visitors. We want customers to feel good about their entire booking experience and come back many times.”

Senior management encouraged these discussions through internal forums, such as the “customer experience debate group” on Facebook’s Workplace collaboration platform. Vismans noted:

People show examples of experiments, which they felt were crossing a line or pushing too much or where we were not fully transparent with customers. We make this a public debate. We know that there is an enormous benefit in having a single metric, conversion. But it’s not perfect. The perfect metric would be loyalty, but it takes years to test and measure for that, to see if customers remain loyal, so we had to find a proxy. If you do proper A/B testing, it will find the most effective way to influence customer behavior. But there is a bigger question: is this model the most sustainable way to grow your business? We are still in the dark ages; the Internet is only 25 years old. This is like we just invented fire. It will take time to fully understand customer behavior. Of course, if someone wants to run a ‘bad experiment,’ they can do it. That’s the price we pay for autonomy and for

the enormous firepower it gives us. But I've not seen anything that was intentionally bad or morally questionable. Like manipulating people to buy a five star hotel room if they can only afford three stars. So I'd rather stay away from policing or ethical review boards. That's not a scalable solution. You'd create a bottleneck and a testing police doesn't make people feel like they're empowered. I would rather have a community that is self-correcting; a self-healing organization.

Process

Booking's teams had a clear mandate to run experimentation at high speed. To fuel the test pipeline, people had to constantly come up with new ideas, user problems and need areas. Ideas came from talking to users, from using the product themselves to book accommodations, or from past experiments. Teams could also ask for surveys, lab tests, or other qualitative research and received input from customer services on pain points and user preferences. There were so many different channels, operating services, and languages to optimize that finding ideas for testing was not a major issue. Each team managed its idea generating process and test pipeline.

Since Booking's introduction of a formal experimentation process in 2014, teams had to start with a testable hypothesis. Vermeer noted, "Before there were no clear rules. It was basically, you think of a product improvement, you test A and B, and see what gets more clicks. And then you implement and move on to the next test. But you can easily get experimentation wrong when things are that unstructured. We now make people write down what problem they are trying to solve, and to formulate the hypothesis they want to test, in form of a falsifiable statement that could logically be proven wrong. It forces everyone to think things through, to no longer just guess but to collect evidence and learn to how solve customer problems."^a

To help people write better hypotheses, Vermeer's group created a template (see **Exhibit 9**). It stated that a good hypothesis begins with describing a theory or belief, often based on prior evidence, of how a certain condition for a specific audience may change a mechanism, or how a change may improve the audience's experience with the product. (In the example of the yellow "book" button, a theory could be about how a change in the button's color to blue helps users find it more easily.) A team should then specify which metrics could be used to falsify the theory, or what behavior would validate a test (e.g., more users hover and click.) And finally it should state how the change would help the business (e.g., generate more bookings).

Product director Geert-Jan Grimberg recalled an example: "Our mobile conversion rates in Arabic countries were lower than elsewhere. But the data doesn't tell you why. Once we dove into the data, it became clear that the mobile site wasn't 'right-to-left proof.' In Arabic you read from right to left, rather than left to right. This insight led to a simple hypothesis: we can help our Arabic travelers by making their mobile booking experience right-to-left-proof. So we designed an experiment that ran for two weeks. The control A was an Arabic version of a mobile website left-to-right. The variant B was the same version right-to-left. A hypothesis often starts with an insight that comes from quantitative and qualitative research. Some sort of abnormality that you try to understand."

To launch an experiment, teams needed to fill in an electronic form which was visible to all. The form asked a name for the experiment, state its purpose (in free words or by picking common pain points to solve from a drop down menu), name the main beneficiaries (customers, partners), cite past

^a Experiments can refute—but not prove—a hypothesis. This important tenet of the scientific method, in an attribution to Albert Einstein, can be paraphrased as "no amount of experimentation can ever prove me right; a single experiment can prove me wrong." A theory is validated if repeated and rigorous experiments fail to refute its predictions.

experiments it was based on, state the area it was changing, state the number of variants (up to 20), and specify which platform it was running on (e.g., desktop). The default system settings followed central standards that were developed over years. Vermeer noted, “We have baked a lot of the new guidelines and standards directly into the tools. Teams can change settings but they better have a good reason, as they can easily be challenged by their colleagues for doing so.” An important variable was the threshold, or p-value, that indicated test success: concluding that “challenger B” performs better than “control A” (see Exhibit 10 for experimentation terminology). There was no perfect threshold since an experiment’s p-value also measured the chance of mistakenly accepting “challenger B” as the winner (false positive). A stricter threshold would result in fewer test wins; in contrast, a more lenient threshold would yield more false positives. At Booking, a test’s p-value had to fall below 0.10 (90% confidence) for most tests to be considered “statistically significant.” The minimum run time of an experiment was two weeks. Carini explained the duration logic:

It gives us the seasonality cycle of one week and two Sundays to correct for any outliers, such as the World Cup final on one Sunday. It also gives us time to see whether there are any unintended consequences. And it ensures that we reach a minimum number of users, ideally over a million unique visitors per variant, which can be achieved with a run time of 2 weeks. We need big sample sizes to see significant results, as we typically test for very small changes. That’s what A/B testing is suited for best, to take an existing product and apply small consecutive improvements, one at a time, to create a better product. Teams who needed longer run times were encouraged to add multiples of a week. Experiments used for critical management decision-making sometimes run five to six weeks. Experiments with smaller samples, such as limiting it to French customers visiting Italy, could run for several months.

Many settings and processes in the creation of an experiment were automated. For example, the platform randomly divided customers into a control group and into one, or several, variant group(s). Randomization helped to prevent systemic bias, introduced consciously or unconsciously, from affecting an experiment, as it evenly spread any remaining (and possibly unknown) potential causes of the outcome between treatment and control groups. Enggist said, “To our operational people in customer service, who are less involved in testing, I often explain this in metaphors. Say you have a stadium full of people. You give half of them vitamin C. They have lots of other things happening to them, but due to randomization those things are evenly spread over all people, so it’s only the vitamin C that will make the difference.”

While filling the electronic form, the system informed teams about similar experiments currently running; e.g., testing the same functionality of the same product page, and those that were waiting to begin. Teams were asked to use this information to adjust or postpone their experiment if there was too much overlap, interactions, or potential for conflict. Designers were encouraged to talk to their peers working on similar topics early on to coordinate their testing efforts. Booking did not formally restrict the number of experiments on the same topic. Vermeer noted, “This has been requested several times but we don’t have restrictions. Nobody owns any particular part of a product; teams are all free to run tests. They can informally agree on sequencing their experiments when they think it makes sense, but they don’t have to.” Booking’s platform could automatically identify and highlight experiments that caused problematic interactions, so teams could stop them. Carini said, “If you change the color of a button to blue and another team changed the background color to blue as well, then no customer can see the call-to-action.”

Once the experiment was running, teams watched it closely for the first few hours and if their primary or secondary metrics tanked quickly, they could stop the test early. Carini added, “Methodologically speaking this is not very good but commercially we can’t afford to keep a test

running for the correct run time and risk burning the business down in two weeks.” Frisby continued, “This is something that we could have automated, as it is in other companies, but we chose to keep it manual. We have wall boards around the office that show the number of bookings per second and when teams see that number dropping, we expect them to make the right decision. It’s easier for people to isolate the causes. Say the World Cup starts and bookings drop significantly because of that, we don’t want to stop an experiment.”

Booking’s platform also ran automatic data quality checks and sent warning messages if something was odd. A blue flag was informational, yellow meant that there may be an issue with reporting, and red meant there had been a report failure. A pink flag, the worst warning also called “the pink box of doom,” meant the underlying data had been found to be invalid. An experiment’s information was visible to everyone at Booking and empty template fields could trigger immediate inquiries by other employees. Isbrucker noted, “I have subscriptions for several email reports. You can have reports for your team’s tests, for certain people, or for experiments that were either positive or negative on some metrics. And we get a daily digest with summaries of all tests, so I can reach out if there is anything I want to challenge or discuss. I set aside about one hour a day to review other experiments, particularly more impactful ones or those with novel approaches. There is a lot of learning in that. Of course you can only look at a subset. Even if you only look at 10% of the 2000 that result in statistical significance, that’s still over 200 tests over about 2 weeks.” Specific reports with lessons learned were shared for any experiment that had caused a major problem or break down.

On average, nine out of ten tests failed: they had no, or a negative, effect on selected metrics. But an experiment that failed was not a failed experiment. Vismans noted that it was often useful to investigate further. “For example, we were sure people cared about the quality of WiFi in their hotel rooms. We tested a feature that displayed WiFi speed on a 1-100 scale and customers did not care. It was only when we showed whether the signal was strong enough to do email or watch Netflix, that customers responded favorably.” At the end of an experiment, the team assessed its result as significant (color green), moderate, moderately awful, or as just awful. Carini noted, “This allows anyone in our organization, engineer or not, to quickly draw conclusions. For most tests we don’t need 100% certainty. We are not in pharmaceuticals saving lives; quite often we just want to know if a blue button is equal or better than a yellow, and there is no cost to change it. For tests with significant costs, like incentivizing customers with a \$20 voucher, you need a higher standard of evidence.” After its assessment, the team decided whether to scale the treatment to a permanent feature, which then became the new baseline. Zoeter explained, “We are okay to go for small, even tiny, improvements, and to quickly add them to our website. Even a 1% improvement in conversion can have a big impact on our bottom line.” Frisby added, “We can be very fast, as teams are the decision-making unit. The experiment’s owner just hits a button and turns on a feature for millions of people. In other places they’d have to take the results to some committee, which would make that decision. When experimentation is done well and you have the right cultural norms, you don’t need those safeguards.”

Booking also ran experiments on its supplier network — its partners — but this came with numerous challenges. For one, sample sizes were much smaller and the business impact was more uneven. Large hotel chains accounted for much higher volume than small properties, which had to be accounted for. Next, decision-making by partners often involved multiple people and complex IT systems. Would test participants’ behavior reflect the organizations they represent? Finally, frequent interactions between partners and Booking’s platform meant that experiments had to be approached with more caution, so partner participants wouldn’t get frustrated with too many changes.

Partner testing ran on Booking’s central platform and had grown to about 200 concurrent experiments. The run time was two weeks, within which 60% to 70% of partners would visit Booking at least once. Again, teams had full autonomy, tests were visible to anyone and weekly digests for all

partner experiments were distributed broadly. However, finding the right metrics was an ongoing debate. The best metric would be long-term partner value, but as with customer loyalty, this was hard to derive from a single test. Short-term metrics, such as “number of rooms added,” were closer to the conversion metric used for customers but metrics such as “rooms sold” were also considered. Grimberg described the challenges: “There are fewer predesigned features available and we need to be more careful with partners. One of our teams worked for a month on a personalized login feature, studying needs, making mockups. In our core, they would have tested more quickly; maybe with a dummy link, just send a ‘create your family account now’ to customers and then say, ‘sorry we are just testing this, thanks for your interest.’” Due to the frequent interaction with partners, Booking was upfront about its experiments. Grimberg continued, “We discuss the changes they noticed. When testing a big change, like modified rates and availability, we may attach a survey to the variant, ‘Welcome to our new look; tell us what you think.’ After the tests, we get calls with mixed reactions; some really like what they saw and then realize that it was gone after two weeks.”

Management

Booking’s senior management felt that a true experimentation organization also required a different leadership style. Vismans explained, “I came from a classic top-down company where founders were certain they knew what customers wanted and made all the decisions. But I found out that most of the times their beliefs were wrong. At Booking, everybody knows that, so leadership is much less glamorous. You give your employees the KPIs and let them run.” Senior leadership sets the mission and strategic goals, which had recently changed from an accommodations focus to building a “global experience platform.” They now had to translate the new strategy to investments and KPIs before employees were “free to run.” Tans added:

Many leaders would not feel comfortable in our environment. You can’t have an ego, thinking that you always know best. If I, as the CEO, say to someone, ‘this is what I want you to do because I think it’s good for our business,’ they would literally look at me and say, ‘okay, that’s fine, we are going to test it and see if you are right’. When Booking’s previous CEO first arrived from the US, he presented a redesigned logo to the staff. People said ‘that’s great; we’ll check it with an experiment.’ He was baffled but had no choice. The experiment would determine if the logo could stay.

Tans saw coaching, culture, and talent management as her primary roles. She spent much of her time on recruitment; the only way to scale quickly was to bring in as many smart people as possible. Once they were at Booking, it was important to coach them. Tans continued:

If I make others successful, then the company will be at its best. In meetings, I sit there to help rather than to say what is right and what is wrong. And if I see a team struggle with a decision, I help them think it through. My role is to create a place where people can do their best work. It’s important to me that people are proud of their time at Booking. They should feel that they made a difference to customers and travelling.

Senior management also made sure that people didn’t experiment for just experimentation’s sake. This required an acknowledgement of A/B testing’s limitations. Isbrucker said, “If you don’t have enough traffic—enough users to get significant results—you should not run A/B tests. Also, if you don’t know what success looks like for your product, can’t define it for your hypothesis, the experiment is not going to help you. And testing will give you the ‘what people are doing,’ not the ‘why’ or ‘how’ they feel; to get that you need qualitative research. Finally, testing only offers limited insights on ‘where’ to go next.”

Experiments were most suitable for incremental innovation. Testing a completely new product was difficult and uncomfortable, as there was no baseline to compare it too. Senior Product Owner Deepak Gulati noted, “When you have a strong culture of experimentation that makes incremental improvements to an existing product, there comes a point when the people who built the original product are gone and new products are not in your DNA any longer. You have become a lean and mean machine for customer conversion, for micro-optimizations, driven by experimentation. But when you want to branch out into new areas, you no longer have people who think big, who know how to do this.” Vismans agreed, “This is a downside to a small-step, data-driven organization. We freeze like a deer in the headlights the moment there is no data, no baseline to test against. In our industry, any internet opportunity that you don’t invest in may become a future threat.”

Booking had learned these lessons the hard way. In 2014 it had launched a first product extension, the standalone brand website Villas.com for vacation rentals. Management had thought that customers would value a clear separation between booking hotels and private properties and had wanted to respond to Home Away, AirBnB, and other entrants. Vismans said, “We had no data to support our intuition and there was no test before we launched it. In the end, nobody used it and we closed it down a few years later. We did learn that there is an enormous benefit in having a big audience to start with. It confirmed the danger of big investments just based on intuition or market assumptions.”

One problem with testing radical innovations was that Booking’s platform wasn’t suitable for limited tests. Everything ran in a live environment. Frisby noted, “Even if I limit the user base, say I expose something that changes business processes to just 5% of users, that still represents tens of thousands of transactions a day. And if you reduce traffic, you reduce the power of an experiment. Sometimes it’s better to start with an outside prototype and use qualitative testing to build confidence.” Gulati added, “The large repercussions if something goes wrong are one of the reasons why we insist on incremental steps when new people come in with their big ideas; the other is that when changing several things at once you can’t isolate the variable that caused the metric to change.”

Vismans felt that A/B testing was no substitute for leadership when it came to strategic decisions, “Our new strategy [to diversify into other travel areas like attractions] makes us invest in businesses with lower margins than hotel booking; we assume something is going to happen in the future that will warrant that investment. It’s all belief-based, we have some data, but there is no data that tells us that we have a high chance of being successful. Such ‘business model innovation’ can only come from leadership, not from product teams focused on incremental innovation. And to protect new businesses from ‘organ rejection,’ it may be best to create a new small organization outside of the core, with a direct link to the leadership and new metrics.”

Ultimately, harnessing the power of online experiments came down to management and culture. Vismans concluded:

A/B testing is a really powerful tool; in our industry you have to embrace it or die. If I had any advice for CEOs, it’s this: large scale testing is not a technical thing; it’s a cultural thing that you need to fully embrace. You need to ask yourself two big questions: How willing are you to be confronted every day by how wrong you are? And how much autonomy are you willing to give to the people who work for you? And if the answer is that you don’t like to be proven wrong and don’t want employees decide the future of your products, it’s not going to work. You will never reap the full benefits of experimentation.

Moving Forward

By December 2017, Carini felt that Booking had become a true experimentation organization:

The progress we've made in infrastructure and methodology, especially in the last two years, is significant. When I joined about five years ago, it was mainly the back-end developers who set up the tests and about 50% of our experiments were probably not rigorous enough. Now we have lowered the barriers for experimentation dramatically; everyone can test virtually for free, including product owners or copywriters. We also lowered the perceived costs; once you have a hypothesis, you can test very quickly. For a simple copy change, for example to move from 'Book' to 'Book now,' you just need one server and within an hour, you're collecting data. If you want to test a copy translation for 43 languages, it takes 24 hours. If you want to track multiple devices, it can be done in one to two days. In other companies this would take much longer because you need to order the test from dedicated specialists, which creates a backlog.

Tans felt that Booking was ready for the next move, "Booking has moved through different phases: first, it was about defining the product, the model, the culture; second, there was a long phase to scale everything and everywhere, and now we are the biggest in the world. But we still have gaps in customers, accommodations, and markets, such family or business travel. Customers still spend too much time on planning, and part of that is friction. 80% of our Amsterdam customers open our email at the beginning of their trip when we ask if they need help. So we broaden our mission to enrich customers' journeys with more products like attractions, which requires new tools, new customer service complexity, and so on."

With the large travel market, Tans was confident there were still many opportunities for Booking to grow but it also faced challenges. "My biggest fear is that we lose our focus on doing what is best for customers, as we get bigger and more internally focused. We may also get disrupted. Just think about Google getting into flights, becoming the media and the advertiser, or the Chinese company Ctrip looking beyond their home market. Or imagine Amazon with their huge customer base suddenly considering doing hotels for less commission. The competition is huge, so we need to keep innovating." Another challenge was the retention of Booking's deep experience and experimentation culture with a large inflow of new employees. Enggist conceded:

Young employees often just say 'I am going to run some experiments,' then determine their primary metric, look if the tool tells them 'yes' or 'no', and then stop. It's only after they have been here longer that they come to a deeper understanding. There is a 'coming of age' that makes you turn 'Booking blue;' when you look at the experimentation tool and start realizing that you stare at more than just a picture of your test. There are all these other experiments interfering with yours and they need to be aggregated to the bigger context of serving customers. It's like a constant state of balancing a system that you contribute to. I know that if I pull one thread of the sweater, the whole sweater could be unravelling, and that the simple change I'm about to make touches on not just one but potentially 15 threads. 'Booking blue' people have internalized this context; it's how we think and work.

The landing page experiment Frisby had reached out to Tans for advice. He said, "Gillian, I am about to launch an experiment and wanted to give you a heads up. So you don't be surprised if the press finds out. I will launch an entirely new homepage; it will be live for 10% of our customers, just in time for Christmas travel." Frisby had shown her a new landing page, which looked utterly unfamiliar (see **Exhibit 11**). It was completely blue, with a small window in the center: "Accommodations, Flights,

Rental Cars.” All content and design elements—pictures, text, buttons, and messages—that Booking spent years optimizing were gone.

To broaden Booking’s portfolio, Frisby wanted to test a very simple Google-like index page that had the same user interface for accommodations, flights, and rental cars. He found it difficult to introduce new products into a landing page design that was optimized for accommodations. He explained, “I removed everything for the first iteration. In A/B testing, we often move in small steps. But for big changes, I do the opposite and test the most ambitious version first. In the best-case scenario, we will be pleasantly surprised. In the worst case, we have behavioral signals that allow us to make more informed choices in future iterations.” Some colleagues argued that too many changes would make it impossible to isolate causal variables. Frisby was confident that behavioral metrics would help to improve future customer experiences. A big issue was the reaction of millions of Booking customers in the treatment group (“B”, the challenger) when they opened the unfamiliar landing page.

Frisby’s hypothesis was that changing the customer’s perception of Booking from an accommodation to a full service travel platform was difficult. So he wanted to see if that change could be accelerated with a brand new website that wasn’t optimized for accommodations. Frisby smiled, “I wrote a 3,000 word essay on this, as I always say: ‘The length of the hypothesis should be relative to the complexity of the experiment.’ I went into a lot of detail on the business ambition, how we express it in the experiment, the performance benefits, the qualitative metrics collected prior to running the experiment, really everything. Had I written a 3-line hypothesis and started the test, I would have spent a month answering questions from our community.”

The experiment had been particularly complex to set up. Frisby had worked on its development for five to six weeks, when other tests often took just a few hours. Booking did not fulfill flight and rental car bookings on its core platform, but handed it over to its partners. That meant customers landed on a Booking.com branded version of kayak.com or rentalcars.com. Frisby also had to create new metrics to figure out how to measure the financial gains or losses for Booking. Of course, with radical experiments like this, he was nervous about novelty effects and other customer biases, “Frequently visited platforms like Google see a negative impact on users quickly, while we have a low frequency product. You travel two or three times a year, so I don’t know when any novelty biases wear off. So we’ll have to run the experiment much longer than the usual two weeks.” Frisby was aware that working on Booking’s homepage was something that he may not have been able to do elsewhere, “If you hear folks from other companies talk about A/B testing, they often distinguish between business areas in which you do and do not A/B test. For some landing pages, like Google’s search, experiments are out of bounds. Nobody can touch them. But we don’t have such constraints. Nothing is really sacred; you can do anything here. As we say, if the test tells you that the header of the website should be pink, then it should be pink. You always follow the test.”

Vermeer was very skeptical. He bet Frisby a bottle of expensive champagne that the test would “tank,” meaning that it would drive down conversion rates and be stopped much earlier than planned. Frisby laughed, “Such speculations are not unfounded. Big experiments can suffer or fail miserably.” He added, “But I really enjoy such experiments, more so than incremental tests. But they require deep technical knowledge and an in-depth understanding of our business and strategy. Many people have been here for less than a year. They may have run 30 to 40 experiments, so it’s better if they stay with incremental testing.”

Exhibit 1 Booking.com's Popular Landing Page

a) View from Europe

The screenshot shows the Booking.com website interface for a search in Amsterdam. The header includes the Booking.com logo, currency (€), and navigation links like 'List Your Property', 'Register', and 'Sign in'. Below the header, there are tabs for 'Accommodations', 'Flights', 'Flight + Hotel', 'Car Rentals', and 'Airport Taxes'. The main content area displays 'Amsterdam' with a brief description and a 'Map View' button. A search sidebar on the left allows filtering by destination, dates, and number of guests. The main results section shows 'Amsterdam: 1,459 properties found' with filters for 'Our Top Picks', 'Distance from...', 'Lowest Price First', 'Stars', and 'Top Reviewed'. Two featured properties are highlighted: 'NH City Centre Amsterdam' (Very Good 8.3) and 'The Student Hotel Amsterdam West' (Very Good 8.0).

b) View from India

The screenshot shows the Booking.com website interface for a search in Mumbai. The header includes the Booking.com logo, currency (US\$), and navigation links like 'List Your Property', 'Register', and 'Sign in'. Below the header, there are tabs for 'Accommodations', 'Flights', 'Trains', 'Car Rentals', and 'Airport Taxes'. The main content area displays 'Mumbai: 674 properties found – including 175 value deals!' with a brief description and a 'Map View' button. A search sidebar on the left allows filtering by destination, dates, and number of guests. The main results section shows 'Mumbai: 674 properties found' with filters for 'Our Top Picks', 'Lowest Price First', 'Review Score & Price', 'Stars', and 'Distance From Downtown'. Two featured properties are highlighted: 'Taj Lands End' (Excellent 8.8) and 'Great Value Today' (Very Good 8.4).

Source: Company website (www.booking.com).

Exhibit 2 The Priceline Group Key Financials (in USD thousands, except per share data)

	2017	2016	2015
Revenues:			
Agency revenues	\$9,714,126	\$7,982,116	\$6,527,898
Merchant revenues	2,133,017	2,048,005	2,082,973
Advertising and other revenues	833,939	712,885	613,116
Total revenues	12,681,082	10,743,006	9,223,987
Cost of revenues	250,537	428,314	632,180
Gross profit	12,430,545	10,314,692	8,591,807
Operating expenses:			
Performance advertising	4,141,771	3,479,287	2,738,218
Brand advertising	391,584	295,698	273,704
Sales and marketing	561,958	435,225	353,221
Personnel, incl. stock-based compensation	1,659,581	1,350,032	1,166,226
General and administrative	585,541	455,909	415,420
Information technology	189,344	142,393	113,617
Depreciation and amortization	362,774	309,139	272,494
Impairment of goodwill	-	940,700	-
Total operating expenses	7,892,553	7,408,379	5,332,900
Operating income	4,537,992	2,906,313	3,258,907
Other income (expense):			
Interest income	157,194	94,946	55,729
Interest expense	(253,976)	(207,900)	160,229
Foreign currency transactions and other	(35,291)	(16,913)	26,087
Impairment of cost-method investments	(7,587)	(63,208)	-
Total other expense	139,670	193,075	130,587
Earnings before income taxes	4,398,322	2,713,238	3,128,320
Income tax expense	2,057,557	578,251	576,960
Net income (loss)	2,340,765	2,134,987	2,551,360
Total assets	25,451,263	19,838,973	17,420,575
Long-term obligations	11,403,707	8,127,895	7,185,796
Total liabilities	14,187,702	9,990,293	8,625,106
Total stockholders' equity	11,260,598	9,820,142	8,795,469
Net cash provided by operating activities	4,662,036	3,983,731	3,203,523
Net cash used in investing activities	(4,202,035)	(3,333,295)	(3,894,527)
Net cash used in financing activities	(78,663)	(1,297)	(831,288)
Effects of exchange rate changes on cash etc.	99,996	(45,203)	(149,131)
Net increase (decrease) in cash etc.	481,334	603,936	(1,671,423)
Cash, cash equival., restrict. cash, start of period	2,082,007	1,478,071	3,149,494
Cash, cash equival., restrict. cash, end of period	2,563,341	2,082,007	1,478,071
Priceline Group Stock Price (Dec. 31)	\$1,737.74	\$1,466.06	\$1,274.95
Priceline Group Market Capitalization (Dec. 31)	\$92,906.5	\$80,246.6	\$61,304.3

Source: Company documents (Group website) and Capital IQ for market capitalization.

Exhibit 3 The Priceline Group Statistical Data (in USD millions, growth data in million units)

	4Q17	3Q17	2Q17	1Q17	4Q16	3Q16	2Q16	1Q16	4Q15
Gross Bookings^a									
Agency	\$15,015	\$18,594	\$17,947	\$18,140	\$12,978	\$15,757	\$15,369	\$14,534	\$10,344
Merchant	2,965	3,168	2,850	2,546	2,134	2,703	2,494	2,119	1,670
Total	17,980	21,762	20,797	20,687	15,112	18,460	17,862	16,653	12,015
Year/Year Growth									
Agency	15.7%	18.0%	16.8%	24.8%	25.5%	22.6%	19.4%	22.1%	15.3%
Merchant	39.0%	17.2%	14.3%	20.2%	27.8%	40.2%	19.1%	13.5%	(0.9%)
Total	19.0%	17.9%	16.4%	24.2%	25.8%	24.9%	19.4%	20.9%	12.7%
Constant currency	14%	16%	19%	27%	28%	26%	21%	26%	24%
Units Sold									
Room Nights	151.5	177.5	170.2	173.9	129.7	149.6	140.7	136.5	99.1
Year/Year Growth	16.8%	18.6%	21.0%	27.4%	31.0%	29.4%	24.4%	30.5%	26.6%
Rental Car Days	14.7	19.0	20.7	18.6	14.0	18.0	18.5	16.2	12.2
Year/Year Growth	5.4%	5.5%	11.7%	15.4%	14.4%	12.5%	7.9%	10.9%	10.6%
Airline Tickets	1.6	1.7	1.8	1.8	1.6	1.9	2.0	1.8	1.7
Year/Year Growth	3.1%	(11.8%)	(8.7%)	(2.1%)	(4.3%)	(2.5%)	(6.6%)	(7.2%)	(2.6%)
Gross Profit									
	\$2,770	\$4,375	\$2,952	\$2,334	\$2,276	\$3,589	\$2,430	\$2,019	\$1,879
Year/Year Growth	21.7%	21.9%	21.5%	15.6%	21.1%	21.8%	16.1%	20.8%	12.2%
Constant currency	17%	19%	24%	17%	24%	23%	18%	27%	23%

Source: Company documents (Group website).

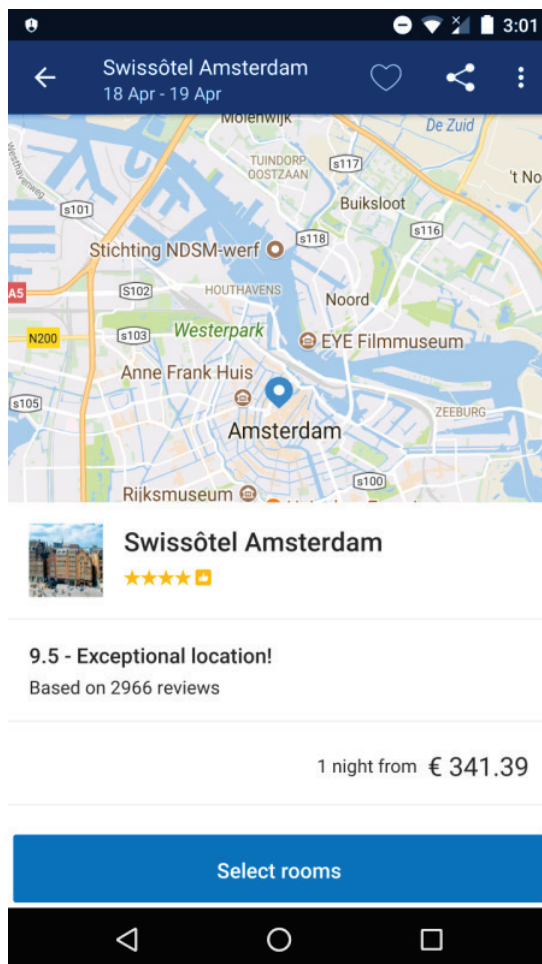
^a Gross bookings is an operating and statistical metric that captures the total \$ value, generally inclusive of taxes and fees, of all travel services booked by our customers, net of cancellations. Amounts may not total due to rounding.

Exhibit 4 Examples of A/B Tests Run By Booking.com

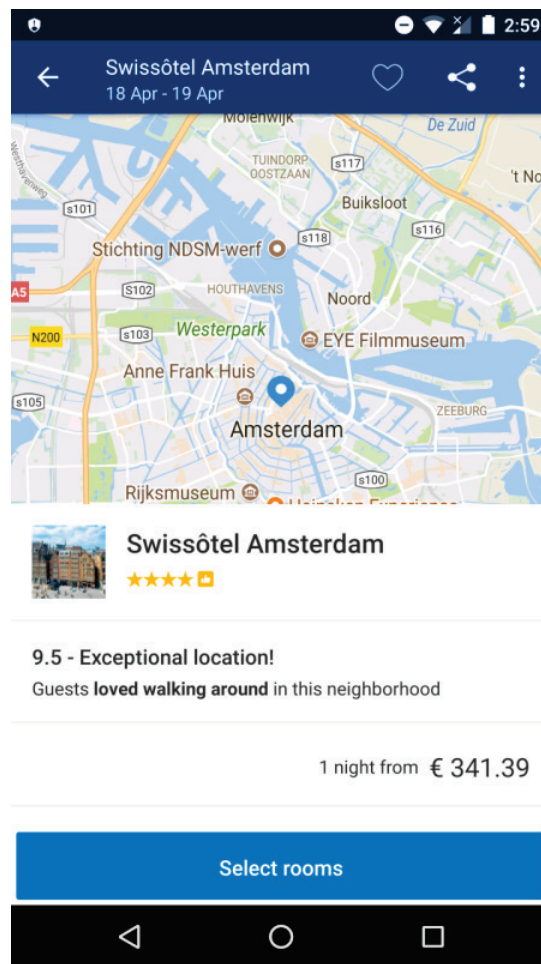
- **Insight:** Research suggested that users care about the area around a property as part of their decision-making process.
- **Hypothesis and A/B Test:** Showing a walkability assessment (i.e., how much guests loved walking around a neighborhood) helps users make better decisions about a property location.

“A”, The Control (Champion)

Shows current practice

**“B”, The Treatment (Challenger)**

Adds walkability assessment (user reviews)



- **Result:** The treatment had no significant impact on key metric; hypothesis not supported and current practice remains champion.

Source: Company interviews.

Exhibit 4 (continued)

- **Insight:** User research suggested that the checkout process could be improved.
- **Hypothesis and A/B Test:** Displaying the checkout date when selecting the number of children improves the user experience (by adding clarity).

“A”, The Control (Champion)

Shows current practice

Rooms
1 ▼

Adults
2 ▼

Children
2 ▼

Ages of children at check-out

4 ▼ 7 ▼

“B”, The Treatment (Challenger)

Adds check-out date above children's ages

Rooms
1 ▼

Adults
2 ▼

Children
2 ▼

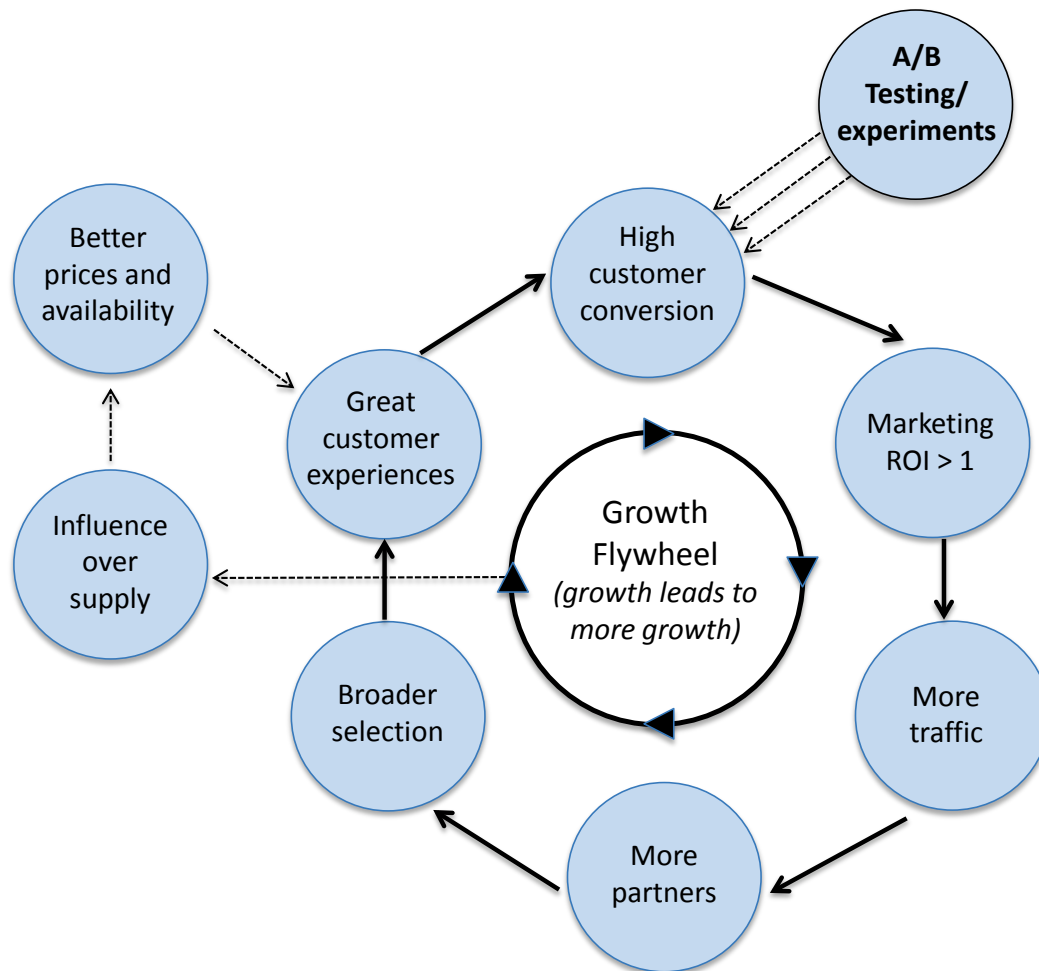
Children's ages on Jul 23, 2016

4 ▼ 7 ▼

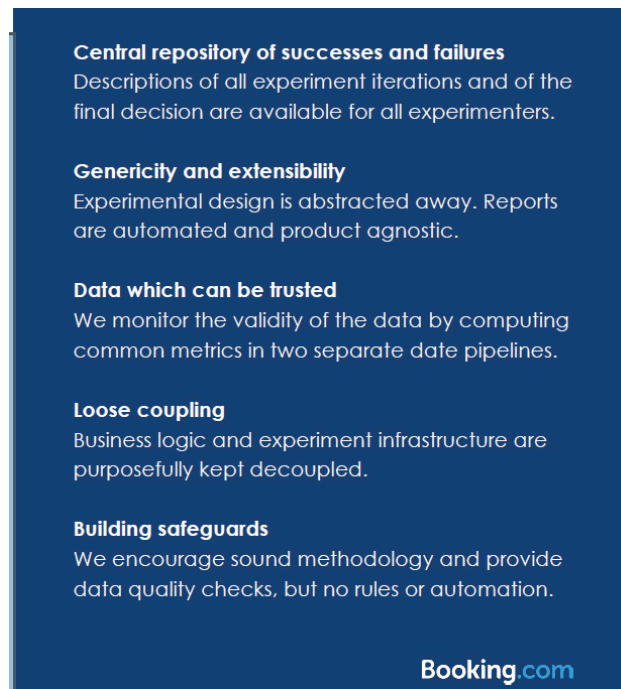
- **Result:** Treatment had a significant positive impact on key metric; hypothesis supported and challenger launched as new champion.

Source: Company interviews.

Exhibit 5 Booking.com's Growth Flywheel



Source: Casewriters (from company interviews).

Exhibit 6 Design Principles of Booking's Experimentation Platform

Source: Company documents.

Exhibit 7 Booking's Core Organization

Department	Size	Comment	Responsibility
Product	180 teams (about 1,200 employees)	Includes marketing development (20 teams, 110 employees) and business units (Booking Home, Payments, Experience)	Drive traffic to website; design and optimize user experience
Partner Services	30 teams (about 190 employees)	Does not include about 8,000 agents who serve customers and partners ^a	Increase accommodations network; design and optimize partner experience
Customer Service	25 teams (about 155 employees)	Does not include about 8,000 agents who serve customers and partners*	Solves booking-related issues
Core Infrastructure	40 teams (about 240 employees)	Does not include IT services	Build, manage and improve technology platform

Source: Company interviews.

^a A part of this workforce is external and seasonal (about 1,500–1,800) to address peak demand.

Exhibit 8 Booking's Shared Values

Value	Explanation
1. We believe in the power of curiosity, experimentation, and continuous learning.	We are genuinely curious and motivated by discovering new possibilities. We are not satisfied by the status quo, nor afraid of failure. Instead we're excited by the constant experimentation required to better understand the needs of our customers and embrace the continuous refinement of our teams, our products, and our processes.
2. We care more about reaching our success together than our individual goals.	We know that teams achieve what individuals cannot, and we thrive on collaboration. We take pride in what we can accomplish together and are happy to put aside our personal ambitions in order to do what needs to be done to succeed as a team.
3. We are humble, open and friendly, knowing our diversity gives us strength.	We know that our true enemy is arrogance, and we remind ourselves everyday how far we still have to go to create the perfect customer experience. It's vital that we're friendly and open. Our natural diversity – in every way imaginable – reflects the diversity of our customers, and the ability to incorporate many viewpoints is critical to our success.
4. We embrace the opportunity to improve and understand that success starts with accountability and ownership.	We each have a part to play and take ownership of our roles with confidence. This means we're not afraid to assume the responsibility we're given, admit when we're wrong or push each other to improve. We are willing to act on behalf of the entire company and know that we succeed only when we both support and challenge.
5. We thrive on change.	Adapting to change is necessary to ensure we are able to respond to evolving customer demands, industry dynamics and high growth. Some people live life at the mercy of change and avoid it at all costs. Others try to cope with change and "just hang in there." At Booking.com, we thrive on it. We believe that rapid change is a driver of opportunity and are excited by what it brings.

Source: Company documents (website <https://workingatbooking.com/about-booking/>).

Exhibit 9 Hypothesis Template & Example

a) Template

Hypothesis template.


Theory	Based on [prior] we believe [condition] for [users] will encourage them to [behavior]
Validation	We will know this when we see [effects] happen to [metrics]
Objective	This will be good for customers, partners and our business because [motivation].

Booking.com

b) Example

Hypothesis Testing.

We observed in user research that some people have difficulty finding the “buy now” button. We suspect this is caused by the low contrast between the font and the background. To solve this user issue, **we will change the button from yellow to blue**. If this solution works, we expect to see more users hover and click, and eventually purchase.



Booking.com

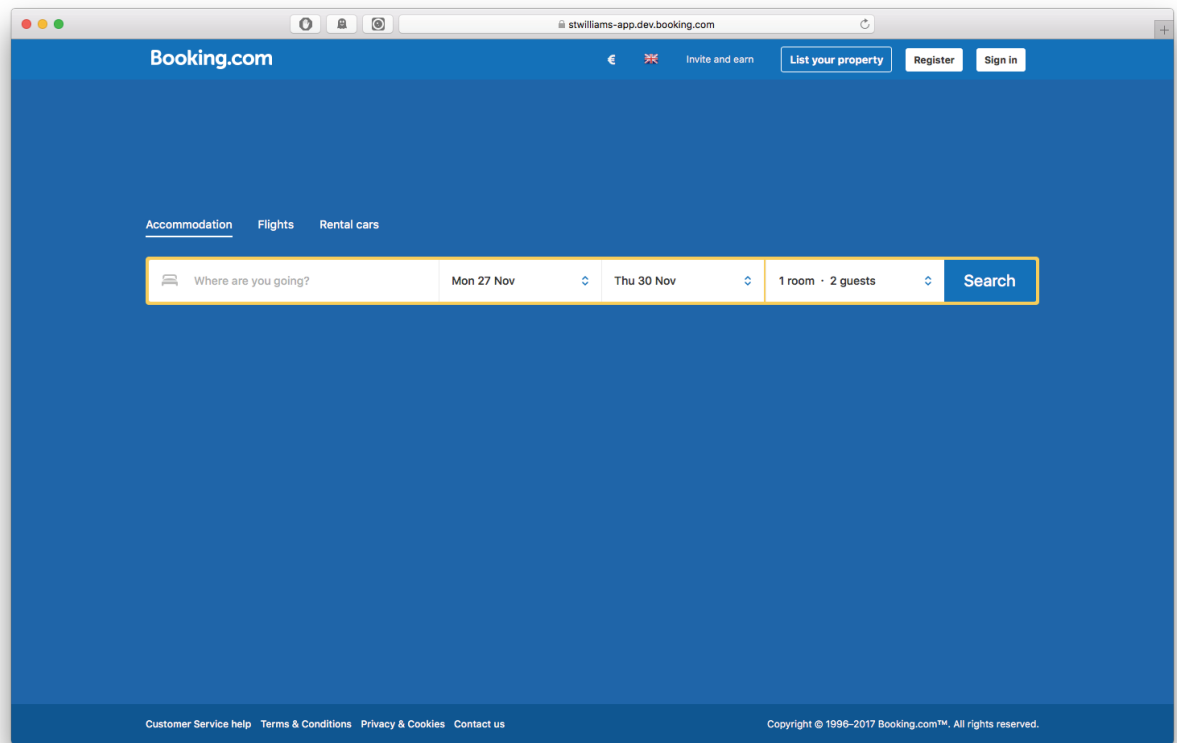
Source: Company documents.

Exhibit 10 Terminology for Offline and Online Experiments

Term	Explanation	Example
Hypothesis	A testable proposition, usually about a treatment's impact on a measurable metric	Opening our stores one hour later ("treatment") will have an impact on daily sales revenue ("metric")
Null Hypothesis	There is no relationship between treatment and metric	Opening our stores one hour later will have <i>no</i> impact on daily sales revenue
Alternative Hypothesis	There is a relationship between treatment and metric	Opening our stores one hour later will have an impact on daily sales revenue
Control	Usually the current practice	No change in store opening hours
Variants	Different treatment levels	One hour late, two hours late, etc.
A/B/n Tests	Users are randomly exposed to control (A) and treatment levels (B/n) for comparison	A current website (A) is compared to variants (B/n) with different colors and their conversion rates are compared
Type 1 Error (False Positive)	Finding a relationship when there is none (<i>rejecting a true null hypothesis</i>)	We conclude that opening stores one hour late has an impact on revenue— <i>even though it has no impact</i>
P-Value	Probability of making a type 1 error (threshold is usually chosen to be 0.05 or 0.10)	At $p=0.10$, there is a 10% chance that we mistakenly conclude that opening stores one hour late impacts revenue
Confidence	Finding no relationship when there is none (<i>failing to reject a true null hypothesis</i>)	Confidence level = $1 - p\text{-value}$. At $p=0.10$, the confidence level is 90%.
Type 2 Error (False Negative)	Finding no relationship when there is one (<i>failing to reject a false null hypothesis</i>)	We conclude that opening stores one hour late has no impact on revenue— <i>even though it does</i>
Power	Probability of finding a relationship when there is one (<i>rejecting a false null hypothesis</i>). Power increases with sample size, magnitude of effect, and significance.	We conclude that opening stores one hour late has an impact on revenue— <i>when it's true</i> . The desired power of an experiment usually falls between 0.8 and 0.95.

Source: Casewriters.

Exhibit 11 The Landing Page Experiment



Source: Company documents.

Endnotes

¹ Statista. "Online travel market – Statistics & Facts." Statista Web site. <https://www.statista.com/topics/2704/online-travel-market/>, accessed July 2018.

² Statista. "Online travel market – Statistics & Facts." Statista Web site. <https://www.statista.com/topics/2704/online-travel-market/>, accessed July 2018.

³ Statista. "Online travel market – Statistics & Facts." Statista Web site. <https://www.statista.com/topics/2704/online-travel-market/>, accessed July 2018.

⁴ The Priceline Group was renamed Booking Holdings in 2018.

⁵ CB Insights. "Where the Big Four Online Travel Agencies – Expedia, TripAdvisor, CTrip, & Priceline – Are Placing Their Bets." CB Insights Research Briefs Web site. <https://www.cbinsights.com/research/expedia-priceline-tripadvisor-ctrip-investments/>, accessed July 2018.

⁶ CB Insights. "Where the Big Four Online Travel Agencies – Expedia, TripAdvisor, CTrip, & Priceline – Are Placing Their Bets." CB Insights Research Briefs Web site. <https://www.cbinsights.com/research/expedia-priceline-tripadvisor-ctrip-investments/>, accessed July 2018.

⁷ Euromonitor International. "Travel Industry and Online Travel Global Overview." Euromonitor International Web site. <http://www.euromonitor.com/travel-industry-and-online-travel-global-overview/report>, accessed July 2018.

⁸ Euromonitor International. "Travel Industry and Online Travel Global Overview." Euromonitor International Web site. <http://www.euromonitor.com/travel-industry-and-online-travel-global-overview/report>, accessed July 2018.

⁹ TechStartups Team. "Google is disrupting the travel industry, killing Expedia, Priceline, and Travelocity with \$14 billion revenue in 2017." *TechStartups*, December 28, 2017. <https://techstartups.com/2017/12/28/google-disrupting-travel-industry-killing-expedia-priceline-travelocity-14-billion-revenue-2017/>, accessed July 2018.

¹⁰ Kim, Tae. "Amazon could disrupt online travel industry next, Morgan Stanley says" *CNBC On the Web*, March 9, 2018. <https://www.cnn.com/2018/03/09/amazon-could-disrupt-online-travel-industry-next-morgan-stanley-says.html>, accessed July 2018.

¹¹ Schaal, Dennis. "The definitive oral history of online travel." *Skift*, July 17, 2016, at <https://skift.com/history-of-online-travel/>, accessed in April 2018.

¹² Schaal, Dennis. "The definitive oral history of online travel." *Skift*, July 17, 2016, at <https://skift.com/history-of-online-travel/>, accessed in April 2018.

¹³ Sorrells, Mitra, "Booking Holdings reveals \$12.7B revenue, goes lukewarm on Airbnb threat." *Phocuswire*, February 28, 2018. <https://www.phocuswire.com/Booking-Holdings-earnings-full-year-2017>, accessed July 2018.

¹⁴ Sorrells, Mitra, "Booking Holdings reveals \$12.7B revenue, goes lukewarm on Airbnb threat." *Phocuswire*, February 28, 2018. <https://www.phocuswire.com/Booking-Holdings-earnings-full-year-2017>, accessed July 2018.

¹⁵ Sorrells, Mitra, "Booking Holdings reveals \$12.7B revenue, goes lukewarm on Airbnb threat." *Phocuswire*, February 28, 2018. <https://www.phocuswire.com/Booking-Holdings-earnings-full-year-2017>, accessed July 2018.

¹⁶ Kohavi, R. and S. Thomke. "The Surprising Power of Online Experiments." *Harvard Business Review* (Sept.-Oct. 2017).

¹⁷ Panyarvudh, Jintana. "Booking a niche in the travel world." *The Nation On the Web*, June 18, 2017. http://www.nationmultimedia.com/news/Startup_and_IT/30318362, accessed July 2018.

¹⁸ Panyarvudh, Jintana. "Booking a niche in the travel world." *The Nation On the Web*, June 18, 2017. http://www.nationmultimedia.com/news/Startup_and_IT/30318362, accessed July 2018.

¹⁹ Pieta, Tomasz. "5 ways to listen to your customers." *Booking.design*, October 24, 2016. <https://booking.design/5-ways-to-listen-to-your-customers-8d06b67702a6>, accessed July 2018.

²⁰ Panyarvudh, Jintana. "Booking a niche in the travel world." *The Nation On the Web*, June 18, 2017. http://www.nationmultimedia.com/news/Startup_and_IT/30318362, accessed July 2018.

²¹ Lukas Vermeer PDF slide presentation: "Democratizing online controlled experiments at Booking.com."