UCLouvain

Faculté des sciences économiques, sociales, politiques et de communication

User Reception of AI-enabled mHealth apps

The Case of Babylon Health

Auteur : Daniela Magalhães Azevedo
Promoteur(s) : Suzanne Kieffer
Lecteur(s) :
Année académique 2020-2021
Master [120] en communication, à finalité spécialisée : gestion de la communication d'organisation et des relations publiques

Acknowledgements

I have to start by thanking my incredible thesis supervisor, Suzanne Kieffer. Your brilliant mind opened my eyes to this new and exciting research field that is UX, but it is the passion that you combine with your expertise that made me fall in love with it. Your wise and thoughtful advice have inspired not only my work, I have grown not only as a professional but also as a human being. So, I thank you, for all your time and your dedication.

To my loving Grigori Epremyan, whose devoted support in trying times and unwavering trust in my abilities makes everything seem possible again, I cannot begin to express how your comforting presence and delicate care have fuelled me with strength during every working hour. From listening to my rambling interrogations, to feeding me during intensive workdays, to proofreading the last draft, you have contributed in ways your modesty prevents you from ever acknowledging. But I do, and I deeply thank you for making every day better than the last.

And to my family and friends, whom I have bored to tears with overdetailed explanations of this thesis for the past half year, I appreciate your patience and your kind queries regarding my work. Know that I intend to pursue this as a professional activity, and I apologize in advance for future long, overly confusing conversations regarding my daily activities. I love you and I am (not) really sorry for this.

Abstract

The proliferation of new technologies empowered by AI offer countless opportunities to improve the current healthcare system. Medical AI already enhances doctors and improves the quality of care provided. Understanding patients' perspectives is paramount to a successful implementation of AI-systems. Despite this, few studies to date have explored the adoptability of these products for patients. Yet, studying patients' reception of mHealth from a user experience (UX) standpoint would enable practitioners to boost adoption rates and thusly accelerate the shift towards smart healthcare. Using an experimental design to compare consumers' reluctances and affinities towards a non-AI product and towards an AI product unveils discrepancies between users' perceptions of product qualities. Furthermore, it highlights issues requiring further research to be properly addressed and ultimately results in a deeper understanding of this crucial, yet neglected, subject.

Table of Contents

1. INTRODUCTION

For decades, Artificial Intelligence (AI) has stimulated the imagination of the scientific community and the public alike. Swift advances in critical elements of this technology have resulted in a proliferation of AI-powered systems. The next decade will see a rapid implementation of 5G which, coupled with recent breakthroughs in quantum computing, could result in AI reaching unforeseen aptitudes. Upcoming breakthroughs will only accelerate the development and multiply application possibilities of this vital technology to appropriately face global economic, social and ecologic crises. Yet, at the time of writing this thesis, little is known regarding the reception of this thrilling yet intimidating technology.

Understanding AI's capabilities as well as its limits is quintessential to progress towards context-mindful models. In the field of medicine, the transition to AI powered solutions makes precision medicine possible and available to a majority of the population on a day-to-day basis. Precision medicine puts the patient's specific characteristics, such as their genes, their habits and their environment, at the centre of the treatment. AI powered systems exponentially increase the accuracy and computability required to produce strategies individually tailored to each patient. This way, healthcare systems can move away from one-size-fits-all types of care that produce ineffective treatment strategies, which sometimes result in untimely deaths.

Professional healthcare workers perspective aside, AI can empower patients and enable them to take control of their healthcare. Recent years have seen a sharp increase in m(obile)Health applications, both in the form of websites as mobile applications, with this very goal in mind. The services offered by mHealth apps range from disease-specific, doctor-prescribed (e.g., supervising glycaemic levels of type 1 and type 2 diabetics, analysing heart health through blood pressure, pulse, cholesterol, glucose levels in blood, tracking medication intake and so on, focused on people bearing a high risk of heart disease) to generalist, occupation-based apps (e.g., step tracker, calories counter, period and related symptoms tracker, and so on). Although the use of most of these apps is centred on personal convenience, the current worsening healthcare crisis deriving from an ageing population with a

decrease in public health funding, urges practitioners and administrators alike to find new solutions to provide affordable care. The COVID-19 pandemic has unveiled the extent of the fragility and exposed many holes in our modern healthcare systems. AI could provide solutions and ease the unrelenting pressure weighing down on our healthcare systems.

However, consumers' general reluctance to engage with AI for sensitive subjects makes implementing AI based solutions complicated. Despite the urgency to repair healthcare systems in anticipation of the next health crisis, few studies have focused on users' perspective on the matter. UX studies offer a considerable arsenal of tools to measure the reception of such applications, howbeit it currently lacks a suiting framework. This thesis aims to shed some, albeit modest, light on this subject and thusly facilitate further research.

The remainder of the paper is structured as follows: Section 2 introduces previous scientific work on this subject. Section 3 presents the research question, hypotheses and the contribution this thesis hopes to make. The employed method is thoroughly explained in Section 4. Section 5 summarizes the obtained results, while Section 6 discusses them and associates the derived findings to previously presented studies. Finally, Section 7 concludes this paper by putting the findings in perspective and making recommendations for future related research.

2. PREVIOUS RESEARCH

2.1. Definition of AI and technologies deriving from AI

Artificial Intelligence (AI) is an umbrella term that refers to technologies that rely on either machine learning, natural language processing, rule-based expert systems, neural networks, deep learning, physical robots or robotic process automation (Davenport, Guha, Grewal & Bressgott, 2020). The idea behind the term Artificial Intelligence is that the program, algorithm, systems and machines demonstrate or exhibit aspects of intelligence and mimic intelligent human behaviour. For Lew and Schumacher (2020), 'AI systems will find their role in the interplay between non-intelligence devices, AI systems (themselves) and human beings' (p. 57).

Not all AI systems are created equal: some are more 'intelligent' than others. AI level of intelligence is best characterized as a continuum between task automation and context awareness (Davenport et al., 2020). Task automation refers to narrow AI that is focused on a specific domain and cannot learn and extend to other domains. Conversely, context awareness is 'a form of intelligence that requires machines and algorithms to "learn how to learn" and extend beyond their initial programming by humans' (Davenport et al., 2020, p. 31). We can already find some AI applications that have surpassed the mere task automation, even though they fall short of context awareness. Nevertheless, the latter form of AI remains distant, and no significant development is expected to occur in the short to medium term. That being said, several technological breakthroughs, such as the implementation of 5G, are expected to galvanize the AI-researchers community and significantly transform fields in which it is applied (Campbell, 2020; Gruson, Helleputte, Rousseau & Gruson, 2019; Ongena, Haan, Yakar & Kwee, 2019 ; Tien, 2017).

2.2. Medical AI

Regardless of levels of intelligence, AI is expected to be the most adopted technology in the future (Davenport et al., 2020) and medicine is one of the fields currently benefiting from and being transformed by AI (Campbell, 2020; Gruson et al. 2019; Lew & Schumacher, 2020). Major forces in this field are technologies based mostly on either Machine Learning (ML) or

Deep Learning (DL). These technologies drive the so-called Fourth Industrial Revolution. Kühl, Mühlthaler and Goutier (2019) describe machine learning as 'computational methods which focus on improving solving problems with computational resources by learning from experience' (p. 354). Deep learning systems are the most able subset of ML. They are based on artificial neural networks (ANNs), they are multi-layered, and are capable of learning without human intervention, they are unsupervised neural networks (Lew & Schumacher, 2020).

These new technologies are highly anticipated and necessary to address several issues within current healthcare systems. Higgins and Madai (2020) explain age-related illnesses are increasing due to ageing populations. Because of this, healthcare systems must imperatively and simultaneously increase the quality of their care and lower their costs to be sustainable. Machine Learning-based technologies have the potential to meet these, at first glance, contradictory goals. Yet, despite the urgent need to develop and implement AI-enabled systems, there is a lack of certified and clinically validated AI products. Lew and Schumacher (2020) best summarise the stakes surrounding AI in medicine, also called medical AI, by saying that 'because it involves human lives, the stakes are naturally high, but so is the potential for high rewards and opportunity' (p. 57).

As of today, medical AI performs as well as humans in some roles (Davenport et al., 2020) and even perform better than humans in specific tasks, such as predicting epidemics (Panesar, 2019). Their predictive abilities have already proven to aid with diagnostics, prognostics and theragnostics (Gruson, Helleputte, & Gruson, 2019). AI is essential to lower misdiagnoses thanks to customised and personalised services based on massive collections of personal medical data (Chung and Park, 2019), harvested from patients with the help of new data collection systems in the form of wearables as well as implantable and ingestible devices (Chowriappa, Dua & Todorov, 2014). Fraser, Coiera and Wong (2018) believe that technologies such as symptoms checkers, when they are well designed, could support diagnosis and improve quality of care, thus augmenting health system performance. However, such technologies need to be thoroughly vetted before hitting the market. A bad design could

reverse the benefits, instead increasing pressure on health systems and putting patients at risk (Fraser et al., 2018).

2.3. Implementing Medical AI

Gruson and al. (2019) predict AI could change the paradigm in healthcare. This new paradigm could be summarized in the 4P's of care: Personalized, Predictive, Preventive and Participatory (Hood and Galas, cited in Heudel, Durand & Blay, 2017). Despite the many advantages of such systems and the accompanying enthusiasm of several medical researchers and practitioners, Lew and Schumacher (2020) present the medical field as a perfect example of overarching issues related to the presence and general role of AI. Most notably, AI's proliferation in medicine multiplied fears of losing human jobs to automation. In fact, Davenport and Kalakota (2019) designate adoption of AI in daily clinical practice as the greatest challenge to AI in healthcare. Some doctors and researchers, like Alexandre (2017), envision a near-future where powerless practitioners will have no choice but to hand over their practice to Silicon Valley and GAFAMs. In Alexandre's fantasy, doctors' capacities would be limited to signing prescriptions they had no say in, thusly hammering the last nail in the profession's coffin.

Yet, it is the case of IBM's Watson that best illustrates the path AI has taken in the past decade: overpromised and underdelivered (Lew & Schumacher, 2020). So far, technical problems have resulted in lack of accuracy and intuition. Zeitoun and Ravaud (2019) respond to Alexander's predictions by presenting two essential elements of medical AI implementation. The first element pertains to the fact doctors must be included in the conception and development of medical AI. Without their expertise, medical AI cannot progress, or even be developed in the first place. In fact, empirical studies have shown that technologies create as many jobs as – and sometimes more than – they destroy. The second element pertains to the realisation that the best results are obtained through a tight knit collaboration between humans and machines. Doctors will most likely remain, at least in the foreseeable future, the ultimate deciders – they are the ones who, in the end, will determine the diagnostic and the prescriptions to administer their patients. Therefore, fears over job loss, such as Alexandre's, have no scientific base,

since, overall, firms are generally moving towards a human enhanced by AI approach anyway, where AI's role is to augment employee's capabilities (Davenport et al., 2020).

Addressing practitioners' fears and reluctances towards AI was the initial aim of this thesis. We were going to address this problem through the prism of implementation of change (Pettigrew & Whipp, 1993), in which individuals' emotional intelligence (Barreiro, & Treglown, 2020) and behaviour (Giraud, Autissier, Johnson, & Moutot, 2013) affect engagement levels (Schaufeli, Bakker & Salanova, 2006) and can diminish the adoptability of new products. Leadership style helps decreasing fear levels while increasing adoption rates (Alilyyani, Wong, & Cummings, 2018). Transformational leadership in particular is known to improve motivation (Altindis, 2011), engagement levels and performance (Buil, Martínez & Matute, 2019), and ultimately improve employee's openness to change (Yue, Men & Ferguson, 2019). However, because of the pandemic, we were unable to go through with the planned experiment, which was set to take place in a hospital. Furthermore, all the firms we contacted afterwards, also refused to be a part of our study.

Despite cited professionals' reservations, Gruson et al. (2019) note a trend towards higher integration levels of data science approaches and decision support systems into healthcare practices. Davenport and Kalakota (2019) expect to see a more extensive use of AI in healthcare within the next 10 years. In medicine, AI works best when it analyses a large corpus of data, especially if it uses Deep Learning systems. By releasing doctors from routine and repetitive tasks, AI powered solutions free doctors' schedules and enable them to spend more time with each patient. This way, doctors and patients can bond and properly discuss complex choices deriving from conditions and treatments (Zeitoun & Ravaud, 2019). This allows doctors to spend more time with patients and use AI to assist them with diagnoses (Lew & Schumacher, 2020; Zeitoun & Ravaud, 2019).

That being said, some healthcare professionals raise legitimate concerns regarding AI. The most notable issues derive from a 'low number of randomized clinical trials to test performance of AI systems, the lack of transparency of information flows within AI applications, the risk of

inequity and discrimination introduced by algorithmic biases, and insufficient regulatory clarity' (Gille, Jobin & Ienca, 2020, p. 2). To these problems, Qiang, Steinfeld and Zimmerman (2019) add that current failure to integrate AI-enabled systems are also due to a lack of contextual integration design. More specifically, they name the inadequate time and place of the meeting between professionals and AI, as well as ineffective form and an overly marked role, as principal culprits for low adoption rates. To be effective, AI-systems need to not only be accurate and performant, but also suited for the cultural and social environment they are put in.

To shift towards precision medicine, AI applications require more data (Panesar, 2019). Indeed, AI, and particularly ML, are able to develop and gain capabilities without explicit human knowledge only as long as they receive enough data (Gruson et al., 2019). Data mining in healthcare faces critical challenges, ranging from collecting to evaluating and interpreting data (Hu, Stiglic & Wang, 2018). Problems such as fragmented or noisy data and biased systems are harmful to the performance of data mining (Hu et al, 2018; Panesar, 2019). To prevent distorted and biased data, data scientists must fill the missing cells in a process referred to as imputation (Lew & Schumacher, 2020). Other solutions can be found in promising studies (e.g., Zahin, Thanh & Hu, 2019) which offer methods to outperform conventional supervised ML with semi-supervised models. Still, overcoming these technical challenges will require time, more data and data scientists.

Ethics need to be at the heart of this discussion, as it pertains to private and sensitive information, such as medical records. As Denecke and Warren (2020) point out, unlike the patient-doctor relationship, confidentiality and privacy are not always guaranteed when using certain types of AI applications. AI apps like chatbots are sometimes present on social media platforms, such as Facebook messenger, and it is unclear how the data collected by these types of apps are being used. This data could be used, sold or marketed by the distributor of the chatbot. In the United States, Sheikh, Sood and Bates (2015) propose to stimulate competition by mandating vendors to open-up their application program interfaces (API) and to review current health information accountability policies to optimize

the balance between data privacy and reuse of data. In the context of scientific research, Gruson et al. (2019) advocate for the use of soft law to balanced perspectives regarding big data and computing, that allow safeguards without handcuffing research efforts to improve healthcare. They argue that an ethical approach to data collection in healthcare requires patients to be informed prior to the use of any AI technology. This approach implies that in order to collect sufficient data to significantly improve AI-systems, not only do healthcare professionals need to adopt AI, but so do their patients.

Currently, many issues pave the way of global AI adoption. One major concern pertains to the so-called 'opaque' operations of AI-systems. Opacity around the underlying algorithms of AI decrease comprehensibility and the lack of transparency negatively impacts the trustworthiness of the system, which hurts the overall user experience (Weitz, Huber, Schiller, André & Schlagowski, 2019). AI is often described as a 'black box'. The output that emerges from data has no traceability to the inputs, because the AI does not understand what it is doing. It is merely creating patterns that are statistical coefficients (Lew & Schumacher, 2020).

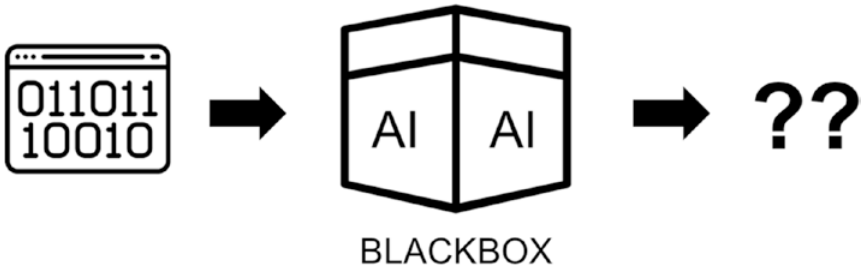


Figure 1 Illustration of information entering the black box (i.e., AI process) and insight emerges on the other side (Lew & Schumacher, 2020, p. 71)

2.4. Medical AI and patients

Aforementioned preoccupations are however not strong enough to slow the quiet departure from an expert-led healthcare approach towards a patient centred, independent and self-sufficient healthcare approach. Since the democratisation of the internet, patients are increasingly looking to (re)gain control over their healthcare choices (Dua & Dua, 2014). This movement is backed by some experts, such as Kish and Topol (2015) who coined the term 'UnPatient' for the new model they advocate. They believe patients

have been subjected to medical paternalism and information asymmetries for far too long and they need to become free to own and use their medical data as they see fit. From their view, democratization of medicine, through the means of digitalisation as well as AI-enabled systems, results in shared control, which in turn provides shared benefits at an exponential rate.

Studies have proven that patients are not only fully capable of possessing and managing their own data, but that it is also beneficial for them to do so, as it increases their sense of well-being as well as enhancing their bond with their physician (Kish & Topol, 2015). Each day, the number of people who turn to websites and social media in search of healthcare information increases (Chowrippa et al., 2014). AI-systems could offer a more suitable alternative to expert-led healthcare, as opposed to simple healthcare information currently available on the internet. Using AI-enabled applications allows users to become more active participants in their own health (Campbell, 2020). In fact, more and more AI-enabled applications enable users to completely bypass encounters with professional healthcare workers (Johnson, 2020). Furthermore, Campbell (2020) notes that these technologies could support the cultural shift towards a patient-centred healthcare paradigm that aims to reduce costs, while at the same time improving patient experience, physician experience and the health of populations.

Nevertheless, as exciting as they may be to some, if these new technologies are to be successfully implemented, it would still require users to heavily change their perception of AI. AI development is currently hampered by customer reservations regarding the technology. In general, customer resistance increases as AI moves towards context awareness (Davenport et al, 2020). Researchers in the early 2000s were already calling for greater attention to consumer resistance regarding technology alternatives (Edison and Gleisser, 2003). The more reservations customers emit, the less likely they are to adopt AI-based products.

Somat (as cited in Distler, Lallemand and Bellet, 2018) describes the process of implementing and adopting a new product as a journey through an acceptability-acceptation-appropriation continuum. The process begins with a subjective evaluation before use (acceptability), then after use

(acceptation) and another one once the product has become a part of daily life (appropriation). User experience can play a pivotal role in the acceptance process (Lew & Schumacher, 2020). Indeed, Sheikh et al. (2015) found that poor usability of current AI-enabled systems undermines efforts to deliver integrated patient-centred care, while Campbell (2020) underlines the importance of usability in regard to wide-spread consumer adoption.

2.5. Adoption of medical AI: a UX perspective

Understanding user's subjective evaluation of interactive systems is at the core of UX studies. To attain a successful user experience, AI-based products are not different from regular products: they need to meet essential elements of utility, usability and aesthetics (Lew & Schumacher, 2020). Thüring and Mahlke (2007) define user experience (UX) as a compound of emotions and perceptions of instrumental and non-instrumental qualities that arise from the user's interaction with a technical device. Rauschenberger, Schrepp, Perez-Cota, Olschner and Thomaschewski (2013) describe the concept of user experience as a combination of aspects of efficiency, aesthetics, attractiveness and joy of use. Both divide UX aspects or qualities into two categories: pragmatic (i.e., instrumental) qualities and hedonic (i.e., non-instrumental) qualities.

Pragmatic qualities refer to the utilitarian aspects of the interactive system, such as its efficiency or functionality. They are thus closely related to the usability and usefulness of a product (Thüring and Mahlke, 2007). On the other hand, hedonic qualities encompass feelings related to the usage of the system, as well as its aesthetics and perceived attractiveness. To borrow Pucillo and Cascini's words (2014, p. 166) 'using a product with specific character in specific situation leads to consequences, such as emotions (e.g., satisfaction, pleasure), explicit evaluations (i.e., judgements of appeal, beauty, goodness), or overt behaviour (i.e., approach, avoidance). *Satisfaction* appears when one is pleased about the confirmation of the prospects of a desirable event, whereas *pleasure* requires no expectations'.

Both Hassenzahl's and Thüring and Mahlke's models agree user emotions come in to play in a mediating role between the two categories. As Norman (2005) explains in his book, aesthetically pleasing products are considered

easier to use, no matter if that is truly the case. He goes on to explain that early impressions of the product affect long-term perception of it. Thüring and Mahlke (2007) summarise the components and the UX interaction in their Components of User Experience (CUE) model (see *Figure 2*). This model suggests that UX has a direct impact on user's behaviours and intent of use.

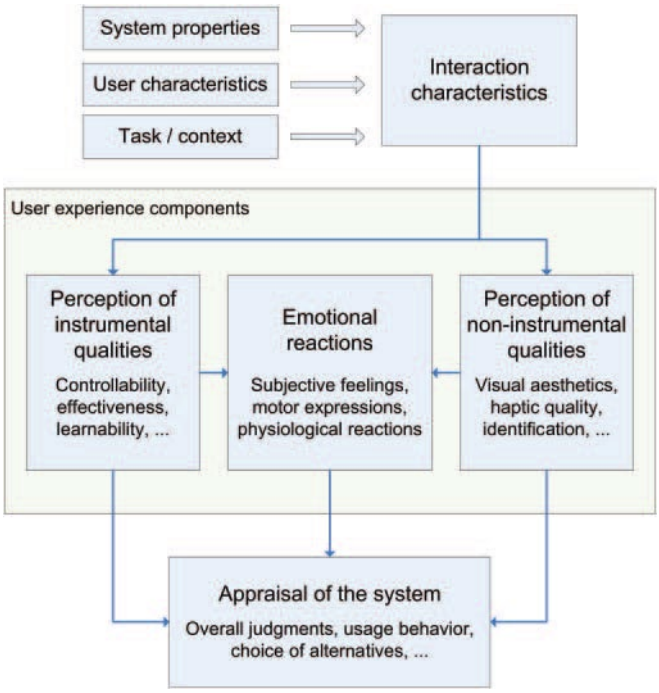


Figure 2 Thüring and Mahlke's CUE model (2007)

Researchers and practitioners have developed UX scales based on models from Hassenzahl (2003) or Thüring and Mahlke (2007) to measure both pragmatic and hedonic aspects of the interaction. Several standardised scales have emerged from these works, such as the meCUE, AttrakDiff and UEQ (User Experience Questionnaire). These tools allow for quantitative and comparative measures to explore relationships between UX constructs in and of themselves or depending on the context of use, system properties or user characteristics. Strong UX metrics can indicate types of effective and emotional responses as well as their extent (Law, van Schaik & Roto, 2014).

However, several problems emerge from current UX measurements and UX methodology. Firstly, standardized UX scales are not suited for all contexts

(Lallemand & Koenig, 2017). Secondly, AI-UX studies often lack in comparability due to the use of different methodologies (Denecke & Warren, 2020) that can stem from different understandings of the same terms, as well as a lack of clear frameworks (Campbell, 2020). Law et al. (2014) go further by arguing that, generally, UX constructs are complex and sometimes impossible to measure in a holistic way (see *Figure 3*). Furthermore, they say that breaking qualities down might not be an ideal solution, because the picture would then become partial. They conclude by saying that to avoid reductionism, UX practitioners should employ quantitative and qualitative methods jointly.

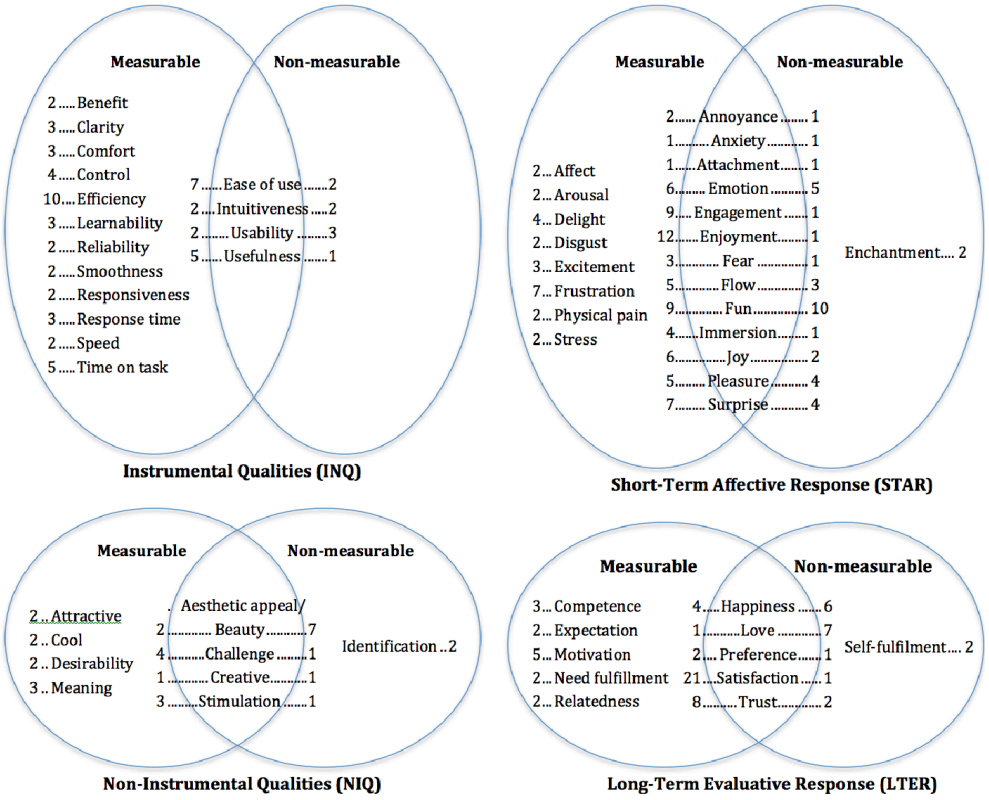


Figure 3 Measurable and non-measurable experiential qualities (Law et al., 2014)

Furthermore, Holmes et al. (2019) conclude that usability testing of a chatbot requires non-traditional methods and that multiple metrics ought to provide a more comprehensive and complete picture of chatbot usability. Holmes et al. (2019) complain that traditional usability testing surveys such as the CUQ, SUS and UEQ do not evaluate all aspects of chatbots.

However, recent improvements in measurement tools might mitigate previous shortcomings in UX measurements. A new well-rounded tool called UEQ+, based on the UEQ, offers a wide variety of new and modular scales. According to its creators, the new UX scales 'can be combined by a UX researcher to build a concrete UX questionnaire adapted to the concrete research question' (Schrepp & Thomaschewski, 2019b). In their accompanying manual, they propose using different scales in accordance with the tested product. For example, for Word Processing applications, they suggest using Dependability, Efficiency and Perspicuity scales. For Info-Websites, they highlight the importance of Content Quality, Trustworthiness of Content and Clarity. However, not all scales are available yet – some are still a work in progress, like Clarity. The conception of these new scales is particularly valuable when trying to understand and to address major problems related to adoption, such as trust. Constructs and scales must be picked carefully though. For example, trustworthiness should not be confused with trust – these are two separate constructs. Trust is highly situational and the trustworthiness of a product does not necessarily lead to a trust relationship (Gille et al., 2020).

The concept of trust is still little understood as we lack clarity on its meaning and dynamics (Gille et al., 2020). Lee and See (2004, p. 54) define trust 'as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability'. Such an attitude requires a balanced calibration to be struck, especially in regard to health-related issues. Overtrust results in a users' trust exceeding system capabilities, while distrust leads users to not fully take advantage of the system's capabilities. Translated to AI, the basis for trust lies in the description of the AI's ability to achieve users' goals. Two fundamental elements define this basis: the focus of what is to be trusted and the information regarding what is to be trusted.

Hoff and Bashir (2015) presented a basic three-layered structure model showing the constructs that influence trust: dispositional trust, situational trust, and learned trust. They define them as follows: 'dispositional trust represents an individual's enduring tendency to trust automation. Situational trust, on the other hand, depends on the specific context of an interaction.

The environment exerts a strong influence on situational trust, but context-dependent variations in an operator's mental state can also alter situational trust. The final layer, learned trust, is based on past experiences relevant to a specific automated system. Learned trust is closely related to situational trust in that it is guided by past experience' (p. 412).

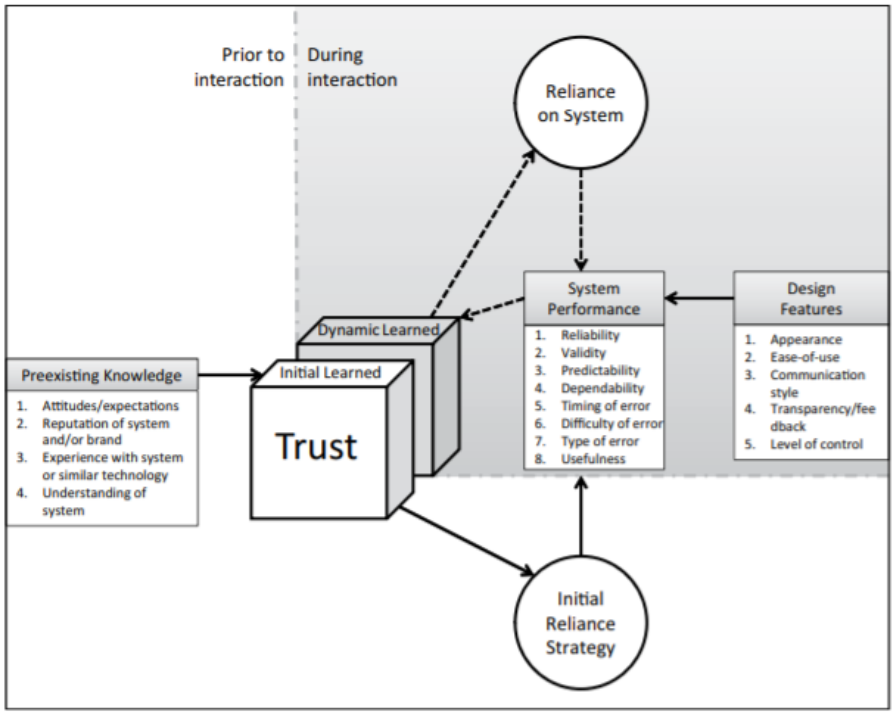


Figure 4 Hoff and Bashir's (2015) model showing factors influencing learned trust. Dotted arrows represent factors that change within the course of a single interaction.

That being said, accuracy, trustworthiness and privacy all play central roles in the user's decision to adopt a new product (Nadarzynski, Miles, Cowie & Ridge, 2019). Indeed, low adoption of AI-enabled systems are linked to low trust rates (Sharan & Romano, 2020). As Lew and Schumacher (2020) explain in their book, trust strongly influences perceived success. One of the main reasons behind AI's failure to build user trust is that their output is often off-point, thus lacking in accuracy. Trust is closely linked with UX pragmatic qualities, in particular usability and utility.

Denecke and Warren (2020) suggest using frameworks that combine users' perception of ease of use with the perception of usefulness of technology to assess acceptance. However, aesthetics cannot be completely disregarded

either due to the *aesthetics usability effect*. This effect refers to users' tendency to perceive the better-looking product as more usable, even when they have the exact same functions and controls (i.e., same utility and usability) (Lew & Schumacher, 2020). The same holds true for user's relationship to the product's brand. This happens because emotions strongly affect the perception of the product (Lew & Schumacher, 2020; Thüring & Mahlke, 2007).

2.6. Major factors influencing adoption

Davenport et al. (2020) summarise the challenges AI adoption faces by discussing four challenges. Firstly, customers hold AI to a higher standard and are therefore less tolerant to error. Secondly, they tend to be less willing to use AI applications for tasks involving subjectivity, intuition or affect, because they believe having those characteristics is necessary to successfully complete the task. Thirdly, consumers are more reluctant to use AI for consequential tasks, such as driving a car as opposed to choosing a movie, because it involves a higher risk. Finally, user characteristics, such as gender, impact adoption rates. In his research, Gustafsod (1998) claims women tend to trust AI less when the risks are perceived to be high. Typically, the last population segment to adopt innovations are older people and/or people with less educational attainment and lower socioeconomic status (Dorsey & Topol, 2020). Incidentally, those are also the people who would benefit the most from these technologies.

Additionally, Araujo, Helberger, Kuikemeier and Vreese (2019) point out that the level of education and programming knowledge also influence associations: people with a lower level of education show a stronger negative attitude towards algorithmic recommendations and respondents with higher levels of programming knowledge show a higher perceived fairness level. In general, they found that people with domain-specific knowledge, belief in equality and online self-efficacy tend to have more positive attitudes about the usefulness, the fairness and risk of decisions made by AI. Apart from demographics, Davenport et al. (2020) say that aspects such as identity also influence adoption. When costumers identify with the domain activity of the AI application, they may be less likely to

adopt AI for that activity. Recent work by Sharan and Romano (2020) show that an individual's personality traits might even override other factors.

Furthermore, the impact of an individual's attitude towards technology exceeds those linked to demographic factors– it is a key factor in the adoption of a wide range of technologies (Edison & Geissler, 2003). Despite the persistent belief that those born after the 80's are digital natives, several empirical studies have shown that technical aptitudes are heterogenous due to variation in age, access to technologies or skills (Bulger, Mayer and Metzger, 2014). Having little technical or digital literacy ultimately influences users' perception of it. According to research dating back to the early 2000s, people behave differently when buying technology. It is in fact a polarising factor – either they like it, or they do not (Edison & Gleisser, 2003).

In their study, Araujo et al. (2019) found that respondents were concerned about potential risks and split about potential usefulness and fairness. Yet, these same respondents evaluated fairness, usefulness and risk for specific decisions in media, public health and justice, of automated decision-making on par and sometimes better than human experts.

A study carried out by Longoni and al. (2019) show that costumers have reservations about AI due to concerns about uniqueness. They found customers generally believe that AI only operates in a standardized manner and is calibrated for the 'average' person. Consequently, AI-systems are perceived as less able to identify and account for costumer's unique characteristics, circumstances and symptoms, leading them to be neglected. Longoni et al. (2019) coined the term *uniqueness neglect* to refer to this concern.

Yet, AI could assess patient's medical records and family history, and even their genome, warn about the disease risk and design unique treatment pathways tailored for them (Panesar, 2019). Same diseases show very different symptoms based on patient's health condition or lifestyle (Chung & Park, 2019) and result in human doctors misdiagnosing patients. Moving care away from hospitals might in fact be safer (Dorsey & Topol, 2020). Indeed, Kish and Topol (2015) explain that in the US alone, 'an estimated 20% of preventable medical errors are due to the lack of immediate access

to health information' (p. 922). Without going as far as taking care away from hospitals, AI-enabled systems could help prevent misdiagnoses by presenting doctors with alternatives diagnoses and information and thusly supporting and enhancing them.

In *Figure 5*, Campbell (2020) summarises the socio-cognitive-technical and cultural factors that research has proven to influence individual's usability of Health Informatics Technology (HIT). These factors are both subjective, such as age or cognitive processing capabilities, and contextual, such as geographic location.

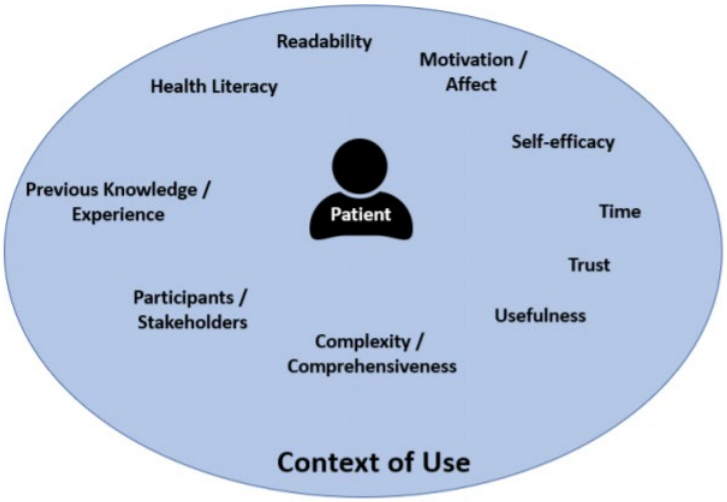


Figure 5 The human factors affecting patient's usability of HIT interacting with the context of use (Campbell, 2020)

2.7. The case of chatbots

Chatbots are predicted to soon become user's preferred interface, over traditional webpages or mobile applications (Cameron et al., 2018). Holmes et al. (2019) argue that interacting with a chatbot is more natural and more intuitive than conventional methods for human-computer interactions. This might explain the rise of chatbots and the current interest in personal digital assistants in this form. A chatbot is an intelligent interactive platform that enables users to interact with AI through a chatting interface (Chung & Parker, 2019) using natural language (written or spoken) aiming to simulate human conversation (Denecke & Warren, 2020). De Gennaro et al. (2020) note that 'chatbots with more humanlike appearance make conversations feel more natural, facilitate building rapport and social connection, as well

as increase perceptions of trustworthiness, familiarity, and intelligence, besides being rated more positively' (p. 3).

However, to decrease complexity, most chatbots opt to restrict user input to selectable predefined items (Denecke & Warren, 2020). We can find chatbots that provide therapeutics or counselling, disease or medication management, screening or collecting medical history or even collecting symptoms to make a triage (Denecke & Warren, 2020). These AI applications have the ability to instantly reach large amounts of users (de Gennaro, Krumhuber & Lucas, 2020), reduce the need of medical staff, as well as assist patients regardless of time and space (Chung & Parker, 2019).

These types of apps are growing in popularity. In the United Kingdom, even before the pandemic, over a million users already preferred to use a chatbot app called Babylon – an app that uses a question-based interaction with users regarding their disease symptoms to establish a diagnosis – rather than contacting their National Health Services (NHS) (Chung & Parker, 2019).

Lew and Schumacher (2020) suggest that AI-enabled assistants should follow Grice's four maxims of communication to be accepted by humans. Grice's (1975) maxims would require AI-apps to be 1) truthful (maxim of quality), 2) as informative as required (maxim of quantity), 3) relevant (maxim of relation) and 4) perspicuous (maxim of manner). Moore et al. (as cited in Holmes et al., 2019) goes further and distinguishes between three types of basic conversation design (Table 1) and four main conversation types (Table 2). For each, Moore et al. (as cited in Cameron et al., 2018) establishes must-adopt practices. For example, for conversational UX, the practices range from progressive disclosure and the breaking down of information into manageable chunks, to the creation of an appropriate and consistent personality.

Principle	Description
Recipient Design	Tailoring conversation to match user's level of understanding
Minimization	Keeping interactions as short and simple as possible
Repair	Recovering from failures and helping understanding (e.g. repeating/rephrasing)

Table 1 Basic conversation design principles (Moore et al., as cited in Holmes et al., 2019)

Conversation Type	Description
Ordinary Conversation	Casual and unrestrained (e.g. small talk)
Service Conversation	Constrained by rules and roles (e.g. customer service agent/customer)
Teaching Conversation	Happens between teacher and student. Teacher probes to test knowledge
Counselling Conversation	Counselee leads the conversation, seeking advice from the counsellor

Table 2 Main conversation types (Moore et al., as cited in Holmes et al., 2019)

In their review of technical metrics used to evaluate healthcare chatbots, Abd-Alrazaq et al. (2020) found that there is no standard method to do so. Most studies use self-administered questionnaires or user interviews. They concentrate on response speed, word error rate, concept error rate, dialogue efficiency, attention estimation and task completion. However, they believe studies should more often include metrics such as appropriateness of responses, comprehensibility or realism. Indeed, chatbot's responses could shape users' experience of such technologies as a whole.

2.8. Simply put

Medical AI represents the future of sustainable, patient-centred form healthcare, with precision medicine at its core. To reach sustainability, on one hand, technologies must evolve and be translated into medical settings and medical culture. On the other hand, patients must adopt these technologies in greater numbers. Only then can we truly alleviate the pressure weighing on our current healthcare systems.

The (non-)adoption of new technologies is a tricky problem that UX researchers have not yet been able to fully conceptualize. They also haven't been able to present and agree upon a single, comprehensible and well-rounded framework. This is because user experience – a subjective evaluation of products – is dependent on a variety of factors, which ought to be tackled from widely different angles. Not only do users' characteristics, such as gender, education level, technology affinity play a major role in the adoption process, but so does the context of use as well as the products'

particular attributes and roles. Popular interfaces such as chatbots require a more in-depth understanding and a precise control over its conversational models and outputs. The more human-like AI products are, the more users tend to appreciate it. Yet, research also tells us that users are more afraid of context-conscient AI, which might seem contradictory, given users' preference of human-like products.

UX researchers and practitioners face many such seemingly contradictory desires. They must consistently consider all aspects affecting user experience and strike a balance among them to create a trust relationship with the user, without which adoption is impossible. But, from an average, non-expert users' perspective, are experiences with AI products truly different from other products?

3. RESEARCH QUESTION

As Longoni et al. (2019) deplore, very little is known regarding patient's receptivity of AI-enabled systems in healthcare. Most studies focus on professionals' perspective and reception, even though patients are the ultimate consumers of these AI systems. Without their seal of approval, we will be unable to meet sustainable healthcare goals. Furthermore, Longoni et al. (2019) note that medical AI enables patients to autonomously collect information and make decisions on their own terms and without being forcefully put under a doctor's guardianship. Therefore, patients will directly drive adoption of AI-enabled systems in healthcare.

Overcoming individual's resistances towards AI-enabled systems is a challenge UX researchers and practitioners must address to achieve better healthcare. Current research already outlined several explanations as to why individuals are particularly reticent towards these types of technologies. However, it is not clear if users are more reticent towards AI-systems than they are towards any kind of technology that could 'replace' human activity. In fact, little is known regarding the influence of branding the product as 'AI'. To our knowledge, recent studies have concentrated their efforts on people's preferences between disclosed AI-systems and human experts (Sharan & Romano, 2020). None have compared the reception of a same product by presenting it (or not) as an AI-system. Yet, understanding if health information provided AI is, in the public's eye, more or less worthy than non-AI-processed information, could help professionals uncover new ways to augment adoption rates. Resolving this issue or, on the contrary, flagging AI-naming would either result in an advancement in the scarce literature on this subject.

To help fill the gap, we try to answer the following question:

> How does knowing Babylon Health is powered by Artificial Intelligence change user reception of the web application?

To answer this research question, we break it down into five testable hypotheses, summarised in Table 3. Each hypothesis contains at least one measurable and comparable components. We design a treatment-based

experiment to verify our hypothesis, which we present in the next section. By analysing and interpreting the results, we hope to shed some needed light on this important yet neglected subject.

	Hypothesis	**Concept: construct**	**References**
H_1	Men and women assess the product differently, including when it is an AI-system.	Gender: female and male	Araujo et al. (2019); Longoni et al. (2019); Gustafsod (1998)
H_2	Knowing Babylon Health operates on a system powered by AI lowers influence perception of content quality and trustworthiness.	Trust: Content Quality, Trustworthiness of Content	Araujo et al. (2019); Hoff & Bashir (2015); Nadarzynski et al. (2019); Schrepp & Thomaschewski (2019a, 2019b);
H_3	Knowing Babylon Health operates on a system powered by AI overall negatively influences the assessment of the pragmatic and hedonic qualities of product.	Hedonic qualities: Attractiveness, Novelty Pragmatic qualities: Efficiency, Perspicuity, Dependability, Usefulness	Hassenzahl (2003); Lew and Schumacher (2020); Thüring and Mahlke (2007); Schrepp & Thomaschewski (2019a, 2019b)
H_4	User's attitude towards technology correlates with the assessment of the product.	Attitude: Technology affinity, technophobia	Edison and Geissler (2003)
H_5	User's attitude towards the device depends on their level of education and expertise.	User characteristics: Educational level, levels of correlated expertise (health or IT)	Araujo et al. (2019); Campbell (2020); Davenport et al. (2020); Longoni et al. (2019)

Table 3 Presentation of research hypothesis, linked concepts and constructs and scientific references

4. METHOD

4.1. Data to collect

Constructs	Methods of collection
Attractiveness	UEQ+, Rank, Interview
Efficiency	UEQ+, Rank, Interview
Perspicuity	UEQ+, Rank, Interview
Dependability	UEQ+, Rank, Interview
Novelty	UEQ+, Rank, Interview
Usefulness	UEQ+, Rank, Interview
Content Quality	UEQ+, Rank, Interview
Trustworthiness of Content	UEQ+, Rank, Interview
Technology Affinity	Attitude towards Technologies Scale
Age	Sociodemographic questionnaire
Gender	Sociodemographic questionnaire
Education level	Sociodemographic questionnaire
Interest in Health	Item 6 of the sociodemographic questionnaire (‘interest’)
Interest in IT	Item 6 of the sociodemographic questionnaire (‘interest’)
Field of work	Item 5 of the sociodemographic questionnaire
Experience with similar app	Interview

Table 4 Constructs and methods of collection

The UEQ is a vetted UX tool that allows easy and valid comparisons between studies as well as between constructs. Furthermore, it is both easy to administer and for participants to understand. The new version of the UEQ, the UEQ+, allows modulation of scales to retrieve specific and desired values.

We chose the first eight UX constructs shown in Table 4 based on previous studies and researchers' recommendations.

In their manual, Schrepp and Thomaschewski (2019b) suggested that dependability, efficiency and perspicuity particularly affect devices that include word processing. They also signalled content quality coupled and trustworthiness of content as qualities to consider when analysing products considered as Info-Websites. Since they offered no recommendation for products such as Babylon, the combination of these two types of devices corresponds best to the requirements.

To this list. we add perceived usefulness because Araujo et al. (2019) strongly suggest perceived usefulness is paramount to adoption of AI. Hoff and Bashar (2015) also interpret usefulness as playing an important role in the 'Learned trust' aspect of trust. To them, overall appearance is another crucial component of user experience of AI. Indeed, the evaluation of user experience would only be very partial, if not outrightly flawed, had we not included hedonic qualities. We settle on the well-rounded and staple of aesthetic, the construct of attractiveness. Lastly, since the product is based on AI, we thought it would be instructive to see how participants perceived the novelty aspect.

We administer a questionnaire to assess attitude towards technology found in Edison and Geissler (2003) to see if it influences user's overall experience of the device as literature suggests. We collect data on participants' gender and educational level for the same reasons. However, we ask participants' age solely for statistical purposes. In our experiment, we target younger adults (20 to 30 years old) as much as possible. Age difference should not significantly impact results.

We ask participants to disclose if they previously engaged with a similar application ('Experience with similar app') because, as Hoff and Bashar (2015) explain, past experiences with an AI-enabled systems may significantly alter the trust formation process, specifically learned trust. Since trust is a crucial component of user experience of AI, we must include these considerations in our study. We also ask about participant's 'Interest in Health', 'Interest in IT' and 'Field of work/study' because participant's expertise on the matter also influences situational trust.

4.2. Methods of collection

Due to the COVID-19 pandemic, data is collected solely through online methods. In the following sections, we first present participant's profile and the tools used to conduct the experiment, including the product containing AI. Then, we present the four methods used to gather data: a two-group experiment, observation, a survey and a semi-structured interview.

Participants

We conduct our research on 20- to 30-year-olds with a minimum of a B2 in English. To participate, they must have access to a stable internet, a computer with a webcam and audio (both microphone and sound are required).

Tools

Zoom

The experiment takes place on Zoom. Zoom is a cloud-based video communication app that allows virtual video and audio set up, such as screen-sharing, recording, conferencing, live chats, and other collaborative capabilities (Antonelli, 2020). Unlike other platforms, no account is needed to attend Zoom meetings, and the platform is compatible with various OS. Its relatively intuitive interface and its simplicity should avoid contaminating the experiment.

Limesurvey

We use Limesurvey to conduct the survey as well as to present the instructions. Limesurvey is a free and open-source online survey application. In our case, all the data we gathered was stocked directly in UCLouvain's servers.

Babylon Health

Since various reports in scientific literature suggest trust in AI-enabled systems often correlates positively with higher degrees of anthropomorphism (Sharan & Romano, 2020), we deliberately chose an AI-enabled system that has no human-like element (behavioural or visual). This way, we do not bias the results of the study by rendering the app more prone

to positive reception. We opted to test our hypothesises on a free-to-use webapp called Babylon Health.

Babylon Health provides various healthcare services through their website and mobile application. Users from the United Kingdom can connect with health care professionals in a manner similar to a text message or hold video messaging consultations, book appointments, track their activity, review their general lifestyle or even receive our prescription and order home blood-test kits (Olivia, 2014; White, 2014). Many of these services are only available to their paying customers.

In this study, we focus solely on their Chatbot Symptom Checker, which is available to every user for free. This chatbot was projected to be an alternative to the UK's National Healthcare Services' (NHS) telephone helpline (O'Hear, 2017). According to Babylon Health's website, their Chatbot is powered by an AI that understands symptoms the users input and provides them with relevant health and triage information. Chung and Parker (2019) confirm that when users enter their symptoms into the app, they receive responses depending on their input. These responses are based on data from a large disease database. At the end of the symptom check, users receive suggestions related to their symptoms. The conversation type (Moore et al., 2017) associated with Babylon is best described as 'Service Conversation'.

Fraser and al. (2018) find impossible to determine how well the Babylon Diagnostic and Triage System – the system behind their Chatbot – would perform on data entered by patients. Other UK researchers found that doctor's diagnoses are correct 77.5% of the time, compared to AI's 90.2% (Donnelly, as cited in Longoni et al., 2019). However, neither the medical validity of Babylon's claims nor the veracity of its diagnoses are discussed in this study.

Observation

We record the call on Zoom to later transcribe the answers, analyse the data closely and to have a record of the answers. The later part prevents data loss in the case of a software malfunction or server shutdown.

Experiment

We use an experimental design (see Figure 6) to test our hypotheses since these designs are well suited to detect differences and cause-effect relationships. Since today most people are unwilling to take a 30-minute survey and respondent's attention and, therefore, answer accuracy declines over time (Lew & Schumacher, 2020; Smyth, 2017), we maximise quality by decreasing question and task complexity as the experiment progresses. We break up the experiment into roughly four manageable parts: (1) a user test (10 minutes), (2) a UX questionnaire and ranking of UX constructs (12 minutes), (3) a short interview (4 minutes) and (4) an attitude towards technologies test followed-up by a sociodemographic questionnaire (5 minutes). When choosing the exact experimental design, we took into account Campbell and Stanley's (1967), as well as Kieffer's (2017), teachings and warnings regarding data validity. We do not include a pre-test to avoid participants guessing the treatment and influencing participants' later answers to the post-test (UX questionnaire). Ideally, we would have two other randomized groups to pre-test. However, due to the modest size and timeframe of this thesis, we are unable to find enough participants to evenly distribute among four groups. For our needs and purposes, this design should be enough: randomized, post-test-only experiment are particularly strong against multiple-group threats, apart from selection-mortality, and they counter single-group threats to internal validity.

	Randomization	*Pre-test*	*Treatment*	*Post-test*
Group 1	**R**		**X**	**O**
Group 2	**R**			**O**

Figure 6 Experimental design employed

Our experiment is structured as follows: after agreeing to participate, respondents start by reading a brief description of Babylon Health in the Limesurvey. Then, they read a situation in which they are asked to use Babylon Health to find a diagnosis. Participants are allowed to read the situation as many times as they wish, while they are using the product and inputting their symptoms. They are instructed to return to the survey once they have reached the result page and have read the results to their liking.

Then, we ask them to share their impressions and subjectively evaluate the product by completing the UEQ+. Once that part is over, participants are tasked with ranking the eight UX qualities by order of importance. We follow-up on their answers with a short semi-structured interview. At the end of the interview, we ask them to stop sharing their screen and to continue the survey. They then answer questions related to their attitude towards technology and they rate their level of technophobia.

In the last part, we ask them to disclose their knowledge of Babylon Health and to input their sociodemographic information.

Treatment

To understand the influence of the term 'Artificial Intelligence' on users' experience of a product, we divide our participants into two groups. The first group receives an introduction of the product that does not mention AI. They are the *No Treatment Group*, also known as the control group.

The second group receives the almost same introduction, except we mention that product is made of AI. They are the *Treatment Group*. In total, we mention the term AI 14 times to this group. We first tell them twice explicitly, once when we present the experiment orally and once when we introduce the app in the written instructions. Then, we tell them 12 times implicitly, by reminding them of this element in the title of the survey and the subtitles above each section of the survey (see Table 5 for an overview). These are the sole differences in treatment.

	N of appearances	**No Treatment Group**	**Treatment group**
Oral description of Babylon	1	You'll be testing Babylon, a Health app	You'll be testing Babylon, a Health app based on artificial intelligence
Written description of Babylon	1	In this experiment we would like to understand what a user experiences when interacting with a product called Babylon Health. Babylon is a British website where you	In this experiment we would like to understand what a user experiences when interacting with a product called Babylon Health. Babylon is a British website where you

		can enter your symptoms and it gives you possible diagnostics. It doesn't replace a doctor, but it gives you an idea of what you could have, and it advises you on the next steps to take.	can enter your symptoms and, based on your input, its artificial intelligence (AI) gives you possible diagnostics. It doesn't replace a doctor, but it gives you an idea of what you could have, and it advises you on the next steps to take.
Survey Title	1	User Experience of Health Apps	User Experience of Health Apps based on Artificial Intelligence
Title above the instructions	2	Babylon Health: a Health App	Babylon Health: a Health App made of Artificial Intelligence
Title above the UEQ+	9	Your personal experience of Babylon Health	Your personal experience of Babylon Health (AI)

Table 5 Differences in Treatment

User test (test)

To minimize interaction and thus avoid experimenter-participant biases, we present a clear written walkthrough, with all instructions presented in the written form. However, at the beginning of the experiment, we very briefly explain how the experiment will unfold and we name the product they will test. We ask participants to share their screen so we can observe their interactions with the product. If they have not already, we also ask them to turn on their webcam so we can see their facials expressions. Then, we give participants a link to the survey.

In the first part of the instructions, we present Babylon Health (see Table 5, 'written description'). Then, we instruct participants to use the product according to the symptoms they are given. We tell them that if a symptom or an act is not mentioned in the description, then it has not happened in that situation. Their goal is to find an appropriate diagnostic using the app. In the following page of the survey, they receive the situation and instruction shown in Table 6.

Situation	It is high season for your seasonal allergies. Your nose has been blocked for the past week, so you have been breathing mostly through your mouth. Your jaw hurts a little and you have been feeling stab-like pain in your right ear that gradually started yesterday. Today, you notice that you are hearing less from that ear. Your ear feels moist, but nothing has come out. When you look at it, everything seems normal. You have no other symptoms.
Instruction	Your ear bothers you a lot, but you don't want to immediately call the doctor. With only this information in mind, you check your symptoms on the website https://www.babylonhealth.com/ask-babylon-chat

Table 6 Situation and instruction given to participants

With only this information in mind, participants are invited to use Babylon Health from their own computers. They are allowed to read the situation as many times as they want throughout the test. To speed up the process, we provided log in and passwords.

Once participants start the symptom check, we do not give them new instructions, but we repeat instructions if they ask. However, if they ask what they should answer or do, we simply tell them to do what they think is relevant to achieve their goal. We translate items at their request.

At the beginning of the test, Babylon proposes to do a covid-check. If participants decide to do a covid-check, we let them go through with it and we ask them why they have done so. Then, we ask them to start a new symptom check and not to choose the covid-check.

Before leaving the Limesurvey to use the app, participants are told that once they reach the result page, they should read it carefully, close the app and return to that page to continue and complete the survey. If, once they reach the result page, participants have forgotten this instruction, are lost and ask for guidance, we reiterate this instruction.

Survey

To avoid acquiescence, which would lower response quality (Smyth, 2017), we use a forced choice format. Whenever possible, we phrase questions to be as little threatening as possible, since there is some evidence that these questions are more effective than direct inquiries (Smyth, 2017).

Most questions are close-ended question because they require less motivation and less skill sets (writing, typing, speaking) from participants (Smyth, 2017). However, close-ended questions leave no place for individual expression. To allow respondents to voice their experience in their own words without burdening them by requiring them to write long answers, we include a short interview during the survey made of open-ended questions.

UX questionnaire (post-test)

A. UEQ+

After interacting with the product, we ask participants to complete the User Experience Questionnaire + (Schrepp & Thomaschewski, 2019a), known as UEQ+. The UEQ+ is a semantic differential with a 7-point Likert-scale for the answers (Schrepp & Thomaschewski, 2019a). Each scale is introduced by a short sentence on top of the items of a scale. This introductory sentence helps participants understand the context for interpretation of the items. Scales have 4 pairs of items, one on each side of the Likert-scale. Pairs of items represent two terms with opposite meanings. The positive term is always on right, and the negative on the left. To further clarify response options, and therefore increase reliability and validity of scales (Smyth, 2017), we label ordinal scales with + and – symbols, as shown in Figure 7.

Example:

	- - -	- -	-	0	+	+ +	+ + +	
annoying	○	⊗	○	○	○	○	○	enjoyable

Figure 7 Example of the UEQ+ scale

To retrieve actionable data for the chosen UX constructs, we used the eight corresponding scales: Attractiveness, Novelty, Dependability, Usefulness, Efficiency, Perspicuity, Content Quality and Trustworthiness of Content.

The meaning of these scales and what they are supposed to measure are presented in Table 7.

Scales	Semantic interpretation
Attractiveness	Overall impression of the product. Do users like or dislike it?
Efficiency	The user has the subjective impression that he or she can achieve the goals related to the usage of the product with minimal effort. The product responds quickly to user actions. The user has the impression that he or she is not forced to enter unnecessary information or to do unnecessary clicks to perform typical tasks
Perspicuity	The user has the subjective impression that it is easy to understand and learn how to use the product.
Dependability	The user has the subjective impression that the product responds predictably and consistently to inputs and commands. The user feels that he or she completely controls the interaction with the product.
Novelty	The user has the impression that the design of the product looks new, fresh and original and catches therefore his or her attention.
Usefulness	The user has the impression that using the product brings him or her advantages. It makes it easier to reach his or her goals, saves time and improves the personal productivity.
Content Quality	The user has the impression that the information provided by the product is actual, well-prepared and easy to understand. It is interesting to read this information.
Trustworthiness of Content	The user has the impression that the information provided by the product is of good quality and reliable. The user has trust in the information provided by the product.

Table 7 Semantic interpretation of the scales as defined by Schrepp and Thomaschewski (2019b) in their manual

We administer the UEQ+ through the same Limesurvey survey. In Table 8, we present our modulated UEQ+ with the introductory sentence participants read to understand the context, as well as the items they rate.

Scales	Introductory sentence	Items
Attractiveness	In my opinion, the product is generally:	annoying / enjoyable bad / good unpleasant / pleasant unfriendly / friendly
Efficiency	To achieve my goals, I consider the product as:	slow / fast inefficient / efficient impractical / practical organized / cluttered
Perspicuity	In my opinion, handling and using the product are:	not understandable / understandable difficult to learn / easy to learn complicated / easy clear / confusing
Dependability	In my opinion, the reactions of the product to my input and command are:	unpredictable / predictable obstructive / supportive not secure / secure does not meet expectations / meets expectations
Novelty	In my opinion, the idea behind the product and its design are:	dull / creative conventional / inventive common / cutting edge conservative / innovative
Usefulness	I consider the possibility of using the product as:	useless / useful not helpful / helpful not beneficial / beneficial not rewarding / rewarding
Content Quality	In my opinion, the information and data provided by the product are:	obsolete / up-to-date not interesting / interesting poorly prepared / well prepared incomprehensible / comprehensible
Trustworthiness of Content	In my opinion, the information and data provided by the product are:	useless / useful implausible / plausible untrustworthy / trustworthy inaccurate / accurate

Table 8 UEQ+ (Schrepp & Thomaschewski, 2019b) modulated for this experiment

B. Ranking of User Experience Qualities

The authors of the UEQ+ ask participants to rate the importance of each construct at the end of each sub-scale. We find this method tedious for

users, so we replace it with a rank-by-importance question at the end (Table 9). When ranking, participants had to keep in mind the specificities of the product (Health App).

Instruction	Items to rank
How important to you are these qualities in a Health App? All your answers must be different and you must rank in order. Double-click or drag-and-drop items in the left list to move them to the right - your highest ranking item should be on the top right, moving through to your lowest ranking item.	Attractiveness Efficiency Perspicuity Dependability Novelty Usefulness Content Quality Trustworthiness of Content

Table 9 User Experience constructs ranking by importance

Constructs are presented in the same order in a 'box' on the left side of the screen. Participants must either drag-and-drop constructs into a separate box, or double click to add them to the last position (screenshots are available in Appendix 9.4.3). Since Perspicuity and Dependability might be considered overly complex or jargon words, especially for none-natives, we provide the following definitions:

Construct	Definition
Perspicuity	Perspicuity is the quality of being clear and easy to understand
Dependability	Dependability is the quality of being trustworthy and reliable

Since respondents generally answer about 4 to 6 items per minute (Ongena et al. 2019), we estimate that respondents should be able to finish the UX questionnaire less than 10 minutes.

Generalities and demographic questionnaire

A. Attitude towards technologies

To measure participants' attitude towards technology, we use a scale provided by Edison and Geissler (2003) from a previously unpublished study. This scale provides 10 statements regarding technology. To facilitate user input, we continued using a 7-point Likert-scale instead of the proposed 5. The goes from Strongly Disagree (on the left) to Strongly Agree (on the right).

N	Items
1	Technology is my friend.
2	I enjoy learning new computer programs and hearing about new technologies.
3	People expect me to know about technology and I don't want to let them down.
4	If I am given an assignment that requires that I learn to use a new program or how to use a machine, I usually succeed.
5	I relate well to technology and machines.
6	I am comfortable learning new technology.
7	I know how to deal with technological malfunctions or problems.
8	Solving a technological problem seems like a fun challenge.
9	I find most technology easy to learn.
10	I feel as up-to-date on technology as my peers.

Table 10 Attitude towards technology questionnaire, found in Edison and Geissler (2003)

We calculate the level of technophobia as follows: mean scores between 1 and 2.49 are considered *Highly Technophobic*, from 2.5 to 3.99 they are *Moderately Technophobic*, from 4 to 5.49 are *Mildly Technophobic* and from 5.5 to 7 are *Not Technophobic*. The result is what we call 'measured technophobia'.

Following Edison and Geissler (2003) suggestion, we also introduce a self-rating question. Participants had to rate their level of technophobia. We call this 'self-reported technophobia'.

Question	Answer form
If 'technophobia' is defined as feeling discomfort about computers or any new technology, which of the following best describes you?	Highly Technophobic Moderately Technophobic Mildly Technophobic Not Technophobic

Table 11 Self-reported level of technophobia, taken from Edison and Geissler (2003)

B. Knowledge about Babylon Health Questionnaire

Participants are then asked to answer the close-ended, single-choice questions presented in Table 12. These questions allow us to see if participants had previous knowledge of the app and if they had noticed Babylon is AI-enabled. The second question targets the No Treatment group, as it gives us an idea of the reliability of our data. We also ask

participants tell us how they think this information influences them to have glimpse of their conscient feelings towards AI.

Questions	Answer format
Before participating in this experiment, had you heard of Babylon Health?	I use it regularly I tried it out I heard of it and/or downloaded it, but never tried I had never heard of it
Have you happened to notice Babylon Health uses Artificial Intelligence?	I noticed at the beginning of the experiment. I noticed while I was testing Babylon. I noticed while I was rating my experience. I noticed after I rated my experience. I noticed now.
How do you think knowing this information affect your opinion of Babylon Health?	I have a more negative opinion. I have a more positive opinion. It does not change my opinion. I don't know.

Table 12 Knowledge of Babylon Health questionnaire

C. Sociodemographic

In the last part of the survey, we ask participants to share key personal information. This information is closely related to other social constructs pertaining to users' characteristics, which, as discussed in the Previous Research section, also influences user experience. By comparing results through various prisms, we broaden the horizon for analysis. Results may also underline biases due to insufficient or over representation of certain populations. This information is particularly useful if this study was to be repeated with other populations.

Questions	Answer format
Age	*Numerical value*
Gender	*Close-ended, single choice:* Male Female Transgender Non-binary I would rather not say
What is the highest degree or level of education you have completed? If currently enrolled, highest degree received.	*Close-ended, single choice:* None Primary schooling Secondary schooling Bachelor's degree

	Master's degree Professional degree Doctorate degree
Are you currently…?	*Close-ended, single choice:* Employed for wages Self-employed Out of work Homemaker Student Military Retired Unable to work
What is your field of work/study?	*Short free input*
What are your interests?	*Close-ended, multiple choice:* Arts and Crafts (i.e., artisanal handicraft or handmade) Intellectual activities (e.g., literature) Outdoor activities (e.g., camping) Cultural activities (e.g., going to the theatre) Indoor activities (e.g., watching TV) Sports (Basketball, Running…) Travel Technology-related activities (e.g., coding) Socialize (e.g., partying) Health & Wellness (e.g., massage) Finances (e.g., comparing prices) Games and puzzles (e.g., video games) Self-educational activities (e.g., reading newspapers) Social and environmental Involvement Religion and spirituality (e.g., pray)

Table 13 Sociodemographic questionnaire

Semi-structured interview (post-test)

Right after participants finished ranking UX qualities, we conduct a short semi-structured interview. Participants' answers to the open-ended questions shown in Table 14 allows us to gather their direct and in-depth commentary on the product. To facilitate communication and obtain a more precise account of their thoughts and impressions, the interview is conducted their native language, French (Table 15).

First, we ask participants to explain the reasoning behind their ranking. Understanding users' priorities is the cornerstone of a good user experience evaluation. To assess the success of the product, we ask participants to

describe their experience. We also ask them if the product met their expectations, to better understand their scoring of the third item of the Dependability scale, in UEQ+ (see Table 8, under 'Dependability'). We also ask them if something worked differently than what they expected in order to identify problems they might have encountered and that impacted their overall experience. Then, to highlight high and low points of their experience, we ask them their favourite and least favourite aspects of the product. Finally, we ask them if they had ever used a similar app before. We estimate that this interview should last about 5 minutes.

N	Questions	Examples (if needed)
1	Why do you rank them this way? / Why are these qualities important to you?	If an app doesn't have appeasing colours, you don't enjoy it as much
2	How would you describe your experience of Babylon Health?	I really enjoyed answering Babylon's questions…
3	Did the product correspond to your expectations?	It was brighter than expected
4	What did you like most about the product?	Colours
5	What did you like less about the product?	Font size
6	Did something work differently than expected?	You were not expecting such precise questions
7	Did you use a similar app before?	

Table 14 Open-ended questions for the interview - English version

N	Questions	Examples (if needed)
1	Pourquoi est-ce que vous les avez hiérarchisés de cette manière ? / Pourquoi est-ce que ces qualités sont importantes pour vous ?	Si les couleurs ne sont pas apaisantes, vous n'appréciez pas l'application
2	Comment décririez-vous votre expérience de Babylon Health ?	Vous avez apprécié répondre aux questions de Babylon
3	Est-ce que le produit correspondait à vos attentes ?	Il était lumineux
4	Qu'est-ce que vous avez le plus apprécié dans ce produit ?	Couleurs
5	Qu'est-ce que vous avez le moins apprécié dans ce produit ?	Taille de la police
6	Est-ce que quelque chose a fonctionné de manière inattendue ?	Vous ne vous attendiez pas à autant de questions
7	Est-ce que vous avez utilité une	

	application similaire auparavant ?	

Table 15 Open-ended questions for the interview - French version

4.3. Methods for analysing collected data

Content analysis

Here, we present only content we consider for the purpose of this thesis and its analysis, even though more information could be gleaned from the data, such as in-depth analysis of users' physiological behaviours during the experiment. Appendix 9.5.7 presents a table overview of all key information. Appendix 9.5.8 contains the transcriptions of the interviews.

English Level Evaluation

We evaluate participants' approximate English level by counting how many vocabulary questions they ask, the type of word they do not understand (i.e., very generic words vs jargon words), how long it takes them to answer a question and how many times they seem lost during the experiment. We also take into account participants' own perception of their English level (self-reported before the experiment started, self-reported during the experiment as well as complaints about not understanding words). We qualify their mastery of the language as 'good' if they have very few questions related to language and if, overall, they seem to understand the instructions. When that is not the case, when participants show many language gaps, but they are still able to finish the experiment and the questionnaire, albeit with some help, we qualify their English skills as 'sufficient'.

Quality of Path Taken

We assess the quality of the path participants take in the application (i.e., if, from the moment they clicked Babylon's website link, they took the right steps to achieve a diagnosis) by measuring the overall accuracy of their inputs given the situation they were presented with, as well as the relevance of the results they obtained. When they respected instructions enough and received a diagnosis related to an ear illness (e.g., otitis, Eustachian tube disfunction, impacted earwax), we rated their path as 'good'. However, if participants started with a non-ear related symptom (e.g., jaw pain, fever,

etc.) and/or they input wrong and significant symptoms (e.g., loss of balance, weight loss, fever, muscle weakness and so on) and therefore either did not obtain a diagnosis or obtain an unrelated one, we assess their path as 'wrong'.

Covid Check

If, at the beginning of the experiment, participants decide to do a covid check, we mark it down. We also ask them why they did so to better understand their reasoning.

UX constructs hierarchy

We rank UX constructs by order of importance based on participants' interview responses. To do this, we count qualities, or a group of qualities, participants believe to be being essential and we report them in a table. We do the same for qualities they consider non-important or least important. We do not count those in-between because they most likely match the written ranking question.

Influence of the term 'AI' on subjective evaluation (self-reported)

After disclosing the moment they knew the app was AI-enabled, participants are asked to give their opinion on the impact of this new information on their overall impression of the product. They can either (1) have a more negative opinion, (2) have a more positive opinion or (3) not change their impression at all. We only report the answers of the No Treatment group, since the Treatment group knew it all along.

Statistical analysis

The raw data obtained is available in a separate SPSS document.

Descriptive measures

In Appendix 9.5.1, we calculate the scale's mean, standard variation, variance and confidence interval of the eight UEQ scales. We present them by treatment group and by population. We also present descriptive statistics by treatment for the entire population (i.e., all 29 participants) in Appendix 9.5.2.

Overview

To get a feeling of the gathered data, we present range, minimum and maximum scores, mean scores and standard error by construct, for the entire population.

Cronbach's alpha

We verify the internal consistency of our data by calculating Cronbach's alpha. We test each measured UX construct, depending on their treatment group and their population, to see if are measuring what they should be. We follow George and Mallery's (2003) tiered approach to interpretation, therefore, considering values ≥ .9 as excellent, ≥ .8 as good, ≥ .7 as acceptable, ≥ .6 as questionable, ≥ .5 as poor, and bellow .5 as unacceptable.

Normal Distribution

To decide which test we should use on the gathered data, we first have to verify if it follows a normal distribution. To do so, we analyse histograms, and we use Shapiro-Wilk's method on our eight chosen UX constructs.

Levene's test of Variance

We also assess the equality of variance of our UEQ+ scales because some tests, such as the t-test and MANOVA, have an assumption of homogeneity of variance. Therefore, before using them, we should verify that our constructs meet the prerequisites.

Independent Sample T-test/Mann-Whitney U-test

Most importantly, we use the Independent Sample t-test or Mann-Whitney U-test to verify our most substantial hypotheses, H_1, H_2 and H_3. We set the bar at $p \leq .05$ to see if 1) gender plays a role in the reception of the product and if 2) knowing the product contains AI influences the user's experience.

Effect Size

We calculate effect size to assess how meaningful and practical statistically (in)significant difference found between genders or between treatment groups are. Cohen's (1988) suggests interpreting the effect size as either small, medium, or large. In his proposed convention, he sets small at 0.2, medium at 0.5 and large at 0.8.

Multivariate Analysis of Variance (MANOVA)

We use a MANOVA to see if 1) mastery of the language, 2) the path participants take in the application (i.e., if they have correctly entered the symptoms and, thus, if the length of the experiment was normal), 3) doing a covid-check, 4) attitude towards technologies (both self-reported technophobia and measured technophobia) or 5) having related interests interact statistically significantly with treatment groups, thus influencing user experience of the product. We also verify if the Gender x Treatment interaction is statistically significant.

Ranking Grid – UEQ+ constructs ranked by importance

We report participants' rank of UEQ+ constructs in a table. We calculate the score of each construct by weighing their position in the rank. We attributed 8 points to the construct positioned in first place thus the construct perceived as being the most important, we then attributed 7 points to the second construct and so on. The construct with the lowest score – thus the construct perceived to be the least important – had a score of 1. The overall ranking is based on the sum of participants' individual scores. We temper the results by nuancing it with participant's justification of their rank, which we obtained through the interview.

Person's Correlation/ Spearman's Correlation

We use Person's Correlation or Spearman's Correlation Matrix to measure the strength of association between our UEQ constructs and the direction of their relationship (positive and negative).

4.4. Protocol

Recruitment

We recruited our participants in two ways. First, we asked students of LCOMU2812 to participate in this study. We introduced them to this study during a Teams videocall, at the end of a class given by Pr. Dr. Suzanne Kieffer. Participation in this study was taken into consideration when grading their final work. Students interested in participating needed to read and sign the consent form (Appendix 9.2.1). Then, they had to book a 30

min slot on Doodle. Doodle automatically generated and sent a Zoom invitation.

Later, we asked acquaintances and strangers alike to participate in this study. We recruited them by posting messages on Facebook, or by reaching out to them personally (see Appendix 9.2.4 and Appendix 9.2.5). Like LCOMU2812 students, participants had to book their slots on Doodle. They then received an e-mail with more instructions (Appendix 9.2.3) and the consent document in a PDF format (Appendix 9.2.2) that they had to fill out and sign. They received no compensation.

Experiment

Before the start of the experiment, participants are required to have read and signed the informed consent document. The experiment starts at the beginning of the Zoom call. Figure 8 presents how the experiment goes.

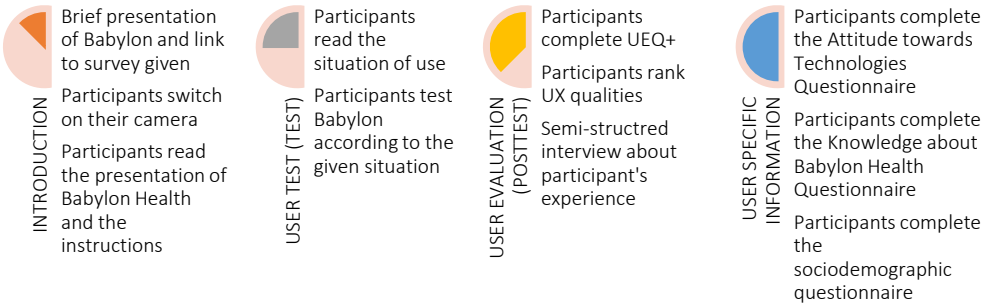


Figure 8 Experiment protocol

4.5. Pilot experience

The method and protocol have greatly evolved since their first iteration, which was tested during the pilot experience. The pilot experience enabled us to see our methodology's shortcomings and receive sufficient feedback to adequately improve our experiment. Since its methodology and results are no longer relevant, we move to Appendix 9.3. In that section, we shortly summarize the methodology and present our findings, interpretations, and improvement decisions.

5. RESULTS

5.1. Corpus data

We started this experiment with 10 males and 12 females. After analysing our results, we decided to widen our range and recruit 8 additional females. Of the 37 people enrolled in our study, we were able to conduct our experiment on 10 males and 20 females. However, we have decided to disregard the results we obtained from one of the female participants. This decision was based on her inability to execute the experiment correctly due to her insufficient English language skills. LCOMU2812 students make for most of our experiment's population, with a group of 10 males and 11 females.

Participants ranged in age from 21 to 34, with a mean age of 24.1. Of the 29 participants, only 2 of them were not students at the time of the experiment (one of which was employed for wages whilst the other one was unemployed), they both finished university during the previous academic year.

Enrolled bachelors and masters university students represent 72.4% of our participants, while the final 27.6% already obtained a master's degree. Most of them (68.96%) are involved, at least to some extent, in communication studies.

	No Treatment		Treatment	
Contingent	**Males**	**Females**	**Males**	**Females**
1st – LCOMU2812	5	6	5	5
2nd – Female	/	10	/	9
3rd – All participants	5	10	5	9

Table 16 Population distribution

We divide our participants into three different sized contingents with an internal division based on 'Treatment' or 'No Treatment' (see Table 16 for an overview). The first contingent consists of (21) LCOMU2812 students exclusively, because we planned the continuity of our experiment based on the analysis of the results we would obtain from this contingent. The second contingent is made up entirely of (19) female test subjects. The no treatment

group contains 10 female test subjects, and the treatment group contains 9 female test subjects. The third contingent consists of (29) all our participants pooled together (LCOMU2812 + female-only). It consists of 5 male and 10 female test subjects for the treatment group, and 5 male and 9 female test subjects for the no treatment group.

Due to the large number of tests we ran, we will only show summary tables in the following section. However, we will reference results from the full tables in the text, these are available in 'Appendices'.

5.2. Content analysis

We estimate that 12 out of the 29 participants only had a sufficient level in the English language. This translated into struggles and the need for assistance to understand instructions and to answer Babylon's questions. Furthermore, 10 participants deviated from the expected path in the application. In some cases they input symptoms which were too broad or incorrect at the beginning of the experiment, which inevitably lead to unrelated questions and results. Other incidents involved the participants inputting invented symptoms – not part of the given instructions - when answering Babylon's questions. Finally, 5 participants started and completed an unnecessary and unasked for covid-check. A full table is available in Appendix 9.5.7.

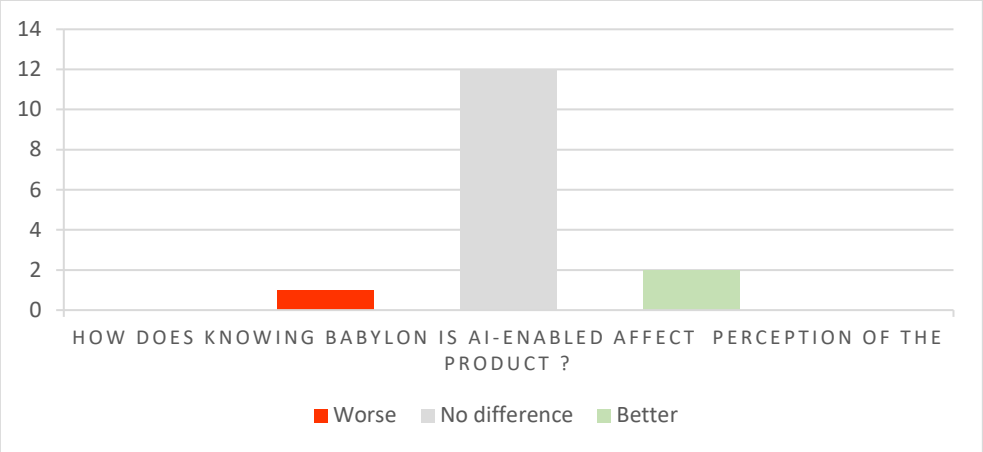


Figure 9 Self-reported perception of AI (No Treatment)

Finally, participants receiving no treatment overwhelmingly reported (80%) their belief that knowing the app is made of AI does not influence their subjective evaluation of the product. Two participants (13.33%) reported

having a better impression of the product after receiving this information and only one participant (6.66%) perceived it to be worse.

5.3. Overview of the results

	Range	Min.	Max.	Mean	
	Statistic	Statistic	Statistic	Statistic	Std. Error
Attractiveness	2.75	4.25	7.00	5.42	.13253
Efficiency	4.00	3.00	7.00	5.42	.18289
Perspicuity	3.00	4.00	7.00	6.18	.15492
Dependability	3.00	4.00	7.00	5.41	.17121
Usefulness	3.50	3.50	7.00	5.33	.18984
Novelty	4.75	2.00	6.75	4.85	.24581
Trustworthiness of Content	3.00	4.00	7.00	5.49	.13843
Content Quality	2.25	4.75	7.00	6.06	.12135

Table 17 Participants' overall evaluation of constructs – range, minimum and maximum scores, mean and standard error

Before analysing, scrutinizing and dissecting the results, let's first take a look at the overall obtained results (Table 17). Participants, regardless of treatment, evaluate the product positively (mean ≥ 5.3 out of 7), except for Novelty, which has a mean score of 4.85. Perspicuity and Content Quality outshine other constructs with a very positive mean score (≥ 6).

Novelty obtains the worst score by far – a mere 2 points – dipping 1 point below the second last worst score, Efficiency, and 1.5 points below Usefulness. Apart from these three constructs, all the other ones reach at least the neutral mark, namely a score of 4 points. Novelty is also the only construct to never obtain a perfect score of 7 points. In general, constructs scores vary from a minimum of 2.25 points in the case of Content Quality, which is relatively small, to a maximum of 4.75 points in the case of Novelty, which either means users evaluate it considerably differently from each other or it indicates the presence of strong outliers.

5.4. Data's Internal consistency

Before we start comparing our groups and statistically analysing our data, we first need to measure the internal consistency of the UEQ+ scale by

using the Cronbach's Alpha (see Table 18) on our gender-mixed 1st contingent – the LCOMU2812 students. This allows us to see the reliability of the scales based on inter-item correlation. Since the experiment was conducted in English on a non-native population, these results are particularly important for a proper interpretation. In this study, though, the results of this test are not necessarily representative of the strength of the UEQ+ scales themselves.

The scales of Attractiveness and Efficiency of our treatment group and the scale of Trustworthiness of Content of our no treatment group only suffer from a slight lack of consistency. However, the Dependability scale of the Treatment group is apparently widely interpreted in unexpected ways by the participants. The results should therefore be interpreted with caution, if not completely disregarded.

	No Treatment Group		Treatment group	
Scale	**Average Corr.**	**Cronbach α**	**Average Corr.**	**Cronbach α**
Attractiveness	0.41	0.74	0.29	0.63
Efficiency	0.59	0.85	0.29	0.62
Perspicuity	0.39	0.72	0.72	0.91
Dependability	0.66	0.89	0.24	0.56
Usefulness	0.46	0.77	0.49	0.79
Novelty	0.86	0.96	0.92	0.98
Content Quality	0.57	0.84	0.42	0.75
Trustworthiness of Content	0.29	0.62	0.63	0.87

Table 18 Cronbach's α of UEQ constructs – LCOMU2812 contingent

The female population – 2nd contingent - provides more consistent results (Table 19). Our results stand overwhelmingly above .7 points, with a maximum of .96. There is however a slight drop on the Attractiveness and Efficiency scales of the treatment group, with only a .67 and .68 respectively, meaning the results obtained for these constructs are questionable.

	No Treatment Group		Treatment group	
Scale	**Average Corr.**	**Cronbach Alpha**	**Average Corr.**	**Cronbach Alpha**
Attractiveness	0.39	0.72	0.34	0.67
Efficiency	0.40	0.73	0.35	0.68
Perspicuity	0.40	0.72	0.84	0.96
Dependability	0.49	0.79	0.43	0.75
Usefulness	0.59	0.85	0.63	0.87
Novelty	0.64	0.88	0.79	0.94
Trustworthiness of Content	0.39	0.72	0.53	0.82
Content Quality	0.46	0.78	0.63	0.87

Table 19 Cronbach's α of UEQ constructs – female contingent

However, when we conducted the same test on our 3rd contingent which combines all participants (Table 20), we see that the Attractiveness and Dependability scales dip below the .6 mark. This means the results are poor and should be analysed with caution, if not discarded altogether. The Efficiency (Treatment) and Trustworthiness of Content scales (no Treatment) also have questionable results. Otherwise, tested constructs have an acceptable α-value, thus making them viable.

	No Treatment Group		Treatment group	
Scale	**Average Corr.**	**Cronbach Alpha**	**Average Corr.**	**Cronbach Alpha**
Attractiveness	0.43	0.75	0.24	0.56
Efficiency	0.50	0.80	0.31	0.64
Perspicuity	0.40	0.73	0.70	0.91
Dependability	0.59	0.85	0.22	0.54
Usefulness	0.55	0.83	0.56	0.84
Novelty	0.70	0.90	0.81	0.95
Trustworthiness of Content	0.48	0.78	0.46	0.77
Content Quality	0.38	0.71	0.57	0.84

Table 20 Cronbach's α of UEQ constructs – all participants

We then proceed to verify that the UEQ+ constructs are normally distributed across all populations (see Appendix 9.5.2 for detailed tables) by using Shapiro-Wilk's test and by looking at the histograms. Table 21 summarises the non-normal distribution and shows that the main distortion of the bell-curve is skewness. When we look closer at the gender-mixed LCOMU2812 contingent (see page 120 for table), we find that male participants were the only ones that did not follow a normal distribution. We used nonparametric tests when measuring nonnormal distribution for tested constructs.

Contingent	Group	Construct	Sig	Distortion
LCOMU2812	Treatment	Dependability	.046	Positive Skew
LCOMU2812	No Treatment	Usefulness	.032	Negative Skew
Female	Treatment	Perspicuity	.006	Negative Skew
All participants	Treatment	Perspicuity	.006	Negative Skew

Table 21 Non normal distribution and error by contingent and by treatment

5.5. Differences between genders (H_1)

We ran an Independent Sample t-test and a Mann-Whitney U-test on our gender-mixed 1st contingent – LCOMU2812 to test our first hypothesis: men and women assess the product differently. In the treatment group, the construct Trustworthiness of Content shows a $t(8) = -3.209$, $p = .012$, thus rejecting the null hypothesis. Females ($M = 6.05$, $SD = 0.64$) graded the construct Trustworthiness of Content more positively than men ($M = 4.95$, $SD = 0.41$) (see *Figure 10* for a visual illustration). We did not detect differences in the no-treatment group, and when testing the groups together, we also did not find differences between genders.

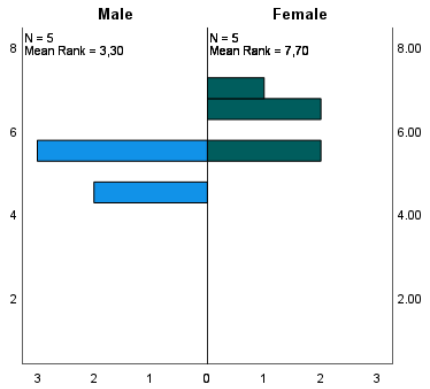


Figure 10 Population Pyramid Frequency for Trustworthiness of Content – LCOMU2812 contingent

To verify the importance of our *p*-value, we then proceeded to calculate Cohen's *d* to analyse effect size on our groups (see Table 22). Following Cohen's (1988) convention, we find that the construct Trustworthiness of Content has a large effect size, d = -2.03, meaning that differences between groups are important as well as statistically significant. Apart from the construct Content Quality, which has a small effect size, all our constructs have medium to large effect sizes (see 'Cohen's *d*' *Table 22* in below). The same does not hold true for our No Treatment group, which generally, has a small effect size. When calculated for the LCOMU2812 – 1st contingent - by treatment, Cohen's *d* shows a mostly small and rarely medium effect size (see Appendix 9.5.5 for all Effect Size tables).

	Standardizer	**Cohen's *d***	**95% Confidence Interval**	
			Lower	Upper
Attractiveness	.66615	-.450	-1.695	.821
Efficiency	.75000	-1.267	-2.617	.146
Perspicuity	1.15244	.477	-.798	1.724
Dependability	.94207	-.796	-2.071	.524
Usefulness	.98900	.404	-.862	1.647
Novelty	1.79496	-.446	-1.690	.826
Content Quality	.54199	-.135	-3.570	-.413
Trustworthiness of Content	.73951	-2.030	-1.372	1.110

Table 22 Effect Size of the LCOMU2812 contingent by treatment group

When we conducted a MANOVA, we found no significant result for the Treatment x Gender interaction, $F(8, 18) = 0.585$, $p = .777$, partial $\eta^2 = .206$. This non-significant value usually means that there is no significant difference between the groups we input (gender and treatment). Indeed, none of the constructs produced a significant result. However, we note that constructs have a very small effect size (apart from the construct Trustworthiness of content, partial $\eta^2 =< .05$), which means the *p*-values have low to no practical importance.

5.6. Differences between treatments (H_2 + H_3)

We then proceeded to test our second hypothesis: people assess the product differently when it is presented as 'based on AI'. Figure 11 illustrates the mean difference with a Confidence Interval of 95% between treatment groups for our gender-mixed 1st contingent – LCOMU2812. The eight constructs are presented side by side and accompanied by their respective standard deviation bars. Appendix 9.5.1 provides the mean, variance, and standard deviation of the UEQ+ constructs by population.

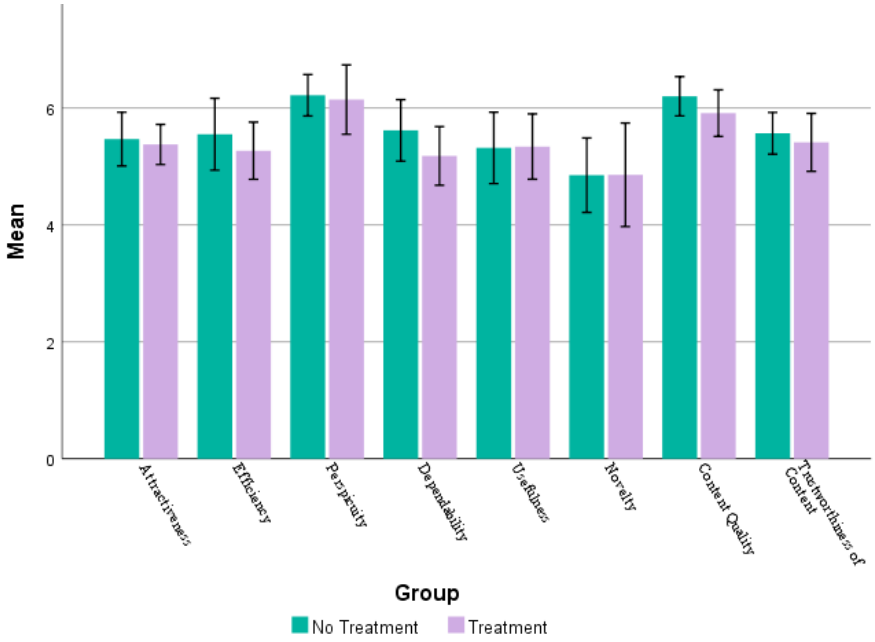


Figure 11 Clustered Bar Mean of UEQ+ constructs with 95% CI by Treatment for the contingent LCOMU2812

Mann-Whitney U-tests and Independent Sample t-tests show what the above illustration suggests: there is no statistical difference between treatment groups in the LCOMU2812 – 1st contingent. We calculate Cohen's *d* and obtain six small effect sizes and two medium ones.

At this point in our research, we decide to expand our population and recruit 8 female participants. We base our decision on three main factors: firstly, our male population is particularly non-normal. Strong statistical tests require a normal distribution to be valid. Normalizing our male population would require more participants than if we continued with our female population. Therefore, creating a larger gender-mixed population would

require more time and work if we are to retrieve actionable data. Secondly, aforementioned research indicated that females tend to react differently from males. Unlike their counterpart, females supposedly tend to evaluate AI-based products more negatively. Hence, expanding the experiment by adding a female-only population would, at least in principle, allow us to gather more significant results and thus answer our main research question in a more efficient manner. Lastly, most constructs showed small effect sizes, meaning there are most likely too many uncontrolled variables influencing the results. So, it is also in an effort to rectify this problem, that we concentrated on a single gender.

Yet, at first, we found no significant results when we administered the Mann-Whitney and Independent t-tests on our female population (see p. 145 for results). Again, this might be due to a generally small effect size (see result Table in p. 158).

To thoroughly verify our second hypothesis, we grouped the treatment group with participants from the no treatment group who noticed that the product was made of AI either at the beginning, during the testing, or while rating their experience. When tested again, we found that $t(17) = 4.128$, $p = .001$, $d = 1.96$ for the Novelty scale (see Figure 12). Participants who noticed it was AI related ($M = 5.56$, $SD = 0.85$) had a more positive opinion of the product than those who did not notice it was AI related ($M = 3.71$, $SD = 1.09$).

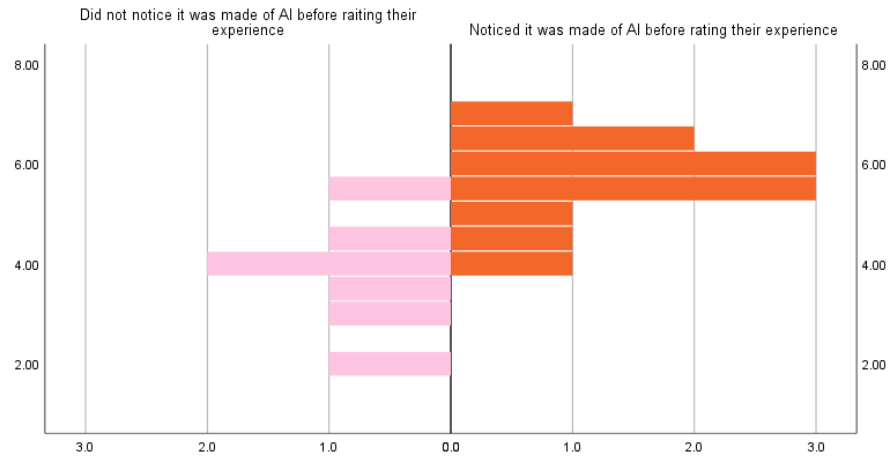


Figure 12 Population Pyramid Frequency for Novelty – female contingent

We found no significant differences between treatment groups when we computed the same tests using all 29 participants results. We came to the exact same realisation when we treated participants' results by moment of knowledge instead of by treatment group. The effect size is overall small, with only two constructs (Dependability and Content Quality) situated closer to a medium effect size.

	Standardizer	Cohen's *d*	95% Confidence Interval	
			Lower	Upper
Attractiveness	.72526	.126	-.604	.854
Efficiency	.99225	.284	-.450	1.014
Perspicuity	.84869	.087	-.643	.815
Dependability	.91108	.481	-.263	1.216
Usefulness	1.04100	-.022	-.750	.707
Novelty	1.34803	-.005	-.734	.723
Content Quality	.64839	.446	-.296	1.180
Trustworthiness of Content	.75484	.207	-.526	.935

Table 23 Effect Size for treatment groups - all participants included

5.7. Other differences (H_4 + H_5)

Finally, we conduct a MANOVA to test the hypothesis that there are one or more mean differences between treatment groups and participants' characteristics (Gender, English language skills, Related Interests, Measured Technophobia, Perceived Technophobia) and specific experience with the product (path taken in the app, performing an unnecessary and unasked for Covid-check). By using a MANOVA, we therefore hope to see if these factors exacerbate a tendency towards either a positive or a negative outcome. However, the multivariate test shows no significant Pillai's Trace or Wilk's lambda, which suggests there is no significant interaction between treatment and participant's characteristics or specific experience with the product (see Appendix 0 for complete tables). Overall, we obtained a large effect size for all interactions, meaning partial η^2 was always $\geq .14$.

5.8. Construct perceived importance

To understand users' priorities when evaluating mHealth apps, we asked them to rank constructs by order of importance. We then created a hierarchy

of constructs by attributing points to the constructs by order of importance. In *Table 24*, we present the calculated mean of the constructs to illustrate how their average positions them in relation to users' priorities (the maximum score is 8, and the minimum score is 1).

Rank	Constructs	Score	Mean
1	Trustworthiness of Content	132	6.95
2	Dependability	118	6.21
3	Usefulness	96	5.05
4	Perspicuity	95	5
5	Efficiency	90	4.74
6	Content Quality	86	4.53
7	Attractivity	38	1.97
8	Novelty	30	1.55

Table 24 Constructs ranked by order of importance, 3rd contingent

Our hierarchy of constructs reveals a clear weight differential when comparing their importance. The first group – the constructs Trustworthiness of Content and Dependability - bundles by far the most important aspects of a Health product. Then, the mean score lowers by one point and we distinguish a second group, containing the constructs Usefulness, Perspicuity, Efficiency and Content Quality, that seem to be judged more or less equally important. Finally, 2.5 points lower than the last construct of that second group, we find the third group consisting of the constructs Attractivity and Novelty.

However, when during a short interview we asked participants to give further explanation concerning their ranking, we found that the construct Content Quality was mentioned as a priority at a higher rate than any other constructs apart from the construct Trustworthiness of Content, which grew its lead even further. The constructs Dependability, Perspicuity and Efficiency maintained a similar level while the construct Usefulness lost 4 ranks. Participants more often expressed their complete disregard for the construct Attractiveness, even though they ranked it higher than the construct Novelty when ranking the constructs by order of importance.

Rank	Constructs	Most important	Least important
1	Trustworthiness of Content	25	0
2	Content Quality	13	0
3	Dependability	6	0
4	Perspicuity	6	0
5	Efficiency	6	0
6	Usefulness	3	0
7	Novelty	0	14
8	Attractiveness	0	18

Table 25 Qualities mentioned during the interview

Several participants suggested during the interview that the Attractiveness and Novelty constructs of the product do not matter to them. Therefore, the results from these qualities ought to interfere with the assessment of other qualities.

5.9. Correlation between UEQ constructs

Despite participants' dismissiveness of the Attractivity and Novelty constructs, the Spearman Correlation Matrix shows (see Table 26) that these constructs correlate positively and significantly with constructs participants judge critical for an mHealth app. For example, the Attractiveness construct correlates significantly with almost every other construct, except for the Trustworthiness Content construct. Indeed, the construct Attractiveness correlates at $p = < .05$ for Dependability, and $p = < .01$ with the constructs Efficiency, Perspicuity, Novelty, Usefulness and Content Quality. The construct Novelty, on the other hand, correlates very significantly ($p = < .01$) with the constructs Attractiveness and Trustworthiness of Content, and it correlates significantly with the constructs Usefulness and Content Quality. Overall, our Spearman's Matrix shows that all constructs significantly and positively correlate with one another.

	Attractiveness	Efficiency	Perspicuity	Dependability	Usefulness	Novelty	Quality of Content	Trustworthiness of Content
Attractiveness	1.000							
Efficiency	.543**	1.000						
Perspicuity	.487**	.298	1.000					
Dependability	.369*	.756**	.456*	1.000				
Usefulness	.598**	.603**	.572**	.573**	1.000			
Novelty	.471**	.350	.135	.199	.414*	1.000		
Content Quality	.476**	.332	.466*	.325	.562**	.374*	1.000	
Trustworthiness of Content	.254	.526**	.172	.502**	.474**	.493**	.618**	1.000
**. Correlation is significant at the 0.01 level (2-tailed).								
*. Correlation is significant at the 0.05 level (2-tailed).								

Table 26 Spearman's Correlation Matrix of all UX measured constructs – 3rd contingent

6. DISCUSSION

Altogether, we have found two statistically significant results with large effect sizes, which are presented in Table 27. Since both of these results have a large effect size and a very significant *p*-value, it is likely they are not the result of a Type I error.

Contingent	Interaction	Constructs	Sig.	Effect Size
LCOMU2812	Treatment group x Gender	Trustworthiness of Content	.012	-2.03
Female	Noticed AI x Did not notice AI	Novelty	.001	1.963

Table 27 Summary of significant findings

These are the sole significant results we have. Indeed, most obtained results are inconclusive. Our computed tests show overwhelmingly small – and sometimes very small – size effects. Therefore, the statistical power of these results is low. We would be committing a Type II error if we were to declare our hypotheses as null based on the results obtained in this thesis. It is possible that if this experiment were to continue, and if we were to increase the size of our sample group, that we would find significant differences between the two treatment groups. If we apply those same premises again, finding absolutely no significant difference between treatment groups is likewise possible. We are simply unable to conclude one or the other with any degree of certainty.

In the following section, we will discuss possible interpretations of the results we obtained, respective to our hypotheses, as well as the adjustments required to obtain conclusive results and the overall validity of the results.

6.1. Hypothesis testing

H_1: difference between genders

The Independent Samples t-test shows female and male participants assessing the products' construct Trustworthiness of Content differently. This finding seems to confirm that which many previous studies reported: when assessing the user experience of a product, gender matters. However, this statement only holds true regarding one of our eight tested constructs. Furthermore, the Multivariate Test found no significant interaction between

Treatment and Gender – both Pillai's Trace and Wilk's lambda were non-significant with a partial η^2 showing large effect size ($\geq .14$).

It is possible that differences between genders might be eroding and expectations standardizing, at least when it comes to one's subjective experience of interactive products. Another possibility is that non-normal distribution, combined with small effect size in the Test of Between-Subject Effects, prevents us from detecting significant differences in the Multivariate Test.

What is interesting in our finding is that women in the treatment group rated the construct Trustworthiness of Content more *positively* than men, whereas we found no difference between men and women in the no-treatment group. Previous work by Gustafsod (1998) found that women tended to be more risk-adverse than men. Applied to this study, Gustafsod's theory should show women rating the construct Trustworthiness of Content *worse* than men in the treatment group. Yet, the opposite is true. This may indicate that gender related association to risk might be outdated and needs to be revisited.

Another possibility is that, unlike Davenport et al. (2020) seem to suggest, the use of AI is not perceived as an increase in risk. In that case, Gustafsod's theory would still hold true. This is plausible if we consider that participants are not entering real symptoms to find a real diagnosis to a real illness, rather they are entertaining a fictional scenario simply because we asked them to. Thusly, they might not perceive the use of Babylon as an increase in risk *for them*, the outcome being imaginary and not translating into palpable consequences cancels the risk perception.

If this is true, then participants may also not feel concerned by uniqueness neglect, as Longoni et al. (2019) conceptualise it, because they only perceive *themselves* as unique. This belief would not necessarily extend to others, even when those 'others' are their hypothetical sick selves. Therefore, in this experiment's context, the principle of uniqueness neglect would not apply.

H_2 + H_3: difference between treatment groups (pragmatic and hedonic qualities)

H_2 and H_3 remain unanswered: we cannot assert that knowing Babylon Health operates on a system powered by AI does not influence subjective evaluation of pragmatic and hedonic qualities. There is only one result presenting serviceable information for interpretation: the Novelty construct in the female population – 2nd contingent.

Female participants scored the Novelty construct more positively when they knew the product was AI-enabled, meaning they tend to find that the design of the product looks newer, fresher, more original and catches more of their attention, when compared to a non-AI-enabled product. They seem to prefer the AI product from this perspective, thus demonstrating there can be positive associations with AI systems and even preference over non-intelligent products. Scientific literature tends to focus heavily on the negative reception of AI and on aspects that require fixing. Because of this bias, little is scientifically known regarding general public's positive views on AI-systems. Yet, this type of information is equally important because products are evaluated as a whole, meaning that positive views can offset the effect of negative views.

H_2 + H_3: importance of the results

The Spearman Matrix shows many correlations between constructs – each construct correlated significantly ($p \leq .05$) with at least four other constructs and very significantly ($p \leq .01$) with at least one other construct. Therefore, all measured constructs influence one another as well as contribute to the overall experience of the product. This supports previous research's conclusion that hedonic and pragmatic qualities both matter when creating user experience. Our results support that all eight constructs need to be considered and addressed to achieve the best outcome. However, this does not mean all constructs are equal in importance. Even though many aspects contribute to the user experience, they do not do so by the same measure, some aspects bear more weight than others.

When we asked participants during the interview to explain their ranking of UX qualities, they mostly focused on aspects linked to constructs

trustworthiness and content quality. Trust was a word that came up particularly often and was treated at length. Aspects related to the constructs content quality and dependability often intertwined with this 'trust' concept. Participants were adamant about the critical role of trust in a mHealth app – without trust in the process as well as the result, they would never adopt said app. During these conversations, the nebulous origin of the data and the app were cited as a cause of concern when judging the accuracy and the legitimacy of the output it provided.

There were slight priority differences in the participants' responses between when they were first asked to establish a hierarchy of constructs and when we came back to that topic during the open interview question. We understand these differences as participants' subconsciously merging the constructs trustworthiness and content quality into a broader concept of trust. Another explanation could be that participants perceive those two constructs as being the pillars on which trust rests.

They frequently neglected aspects linked to the constructs Efficiency, Perspicuity, Dependability and Usefulness. The constructs Novelty and Attractiveness were consistently cited as the counter example of what is considered important in a Health App. In fact, they often portrayed themselves as being very forgiving towards a product's attractiveness and novelty aspects, on the condition that important criteria are met. Although Norman (2005) disproved this particular claim, we believe users' priorities should nonetheless be taken into account in a significant manner when laying the foundations of the UX guidelines, granted their claims are put into perspective and not taken as the end-all be-all of UX design.

H_4 + H_5: difference between users' attitude towards technology, education level and related expertise

We found no significant with large effect size results suggesting measured or perceived technophobia's interaction with treatment leads to differences in users' subjective evaluation. There can be different explanations for this lack of significant results. If the null hypotheses are true, the lack of correlation between UX qualities and attitude towards technology might be due to the employed scale. This scale only addresses the positive attitude

component, whereas we also need to measure aversion towards technology to get a fuller picture (Edison and Geissler, 2003). To observe a significant difference, we might need to rely on a different, more complete measurement tool.

Another possibility could be that we found no significant interaction between attitude towards technology and treatment because our pool of participants does not correspond to what Dorsey and Topol (2020) would identify as typically resistant to adopting new technologies. Indeed, participants in this study are young (M = 24.1 years old) and they have already achieved high levels of education. As a matter of fact, when participants in the No Treatment group were asked if knowing the product is AI-enabled changes their perception of the product, most of them claimed that they would not evaluate it differently. Even though these responses on their own do not answer our research question, they are definitely interesting, because they indicate a low level of (perceived) reservation towards AI. To verify this hypothesis, the same experiment would have to be conducted on older populations and the results of both studies would then have to be compared.

Nevertheless, the non-significance of the results combined with participants' self-reported evaluation being unphased by the presence of AI might also indicate that this generation does not care as much about the term 'AI'. Previous findings regarding users' reservations towards AI would then still be valid and merely not applicable to young and educated populations. There might still be a difference in subjective evaluation in older or less educated populations. Another way to look at it is that there is no statistically significant difference in subjective evaluation between AI systems and non-AI-systems because users have equal reservations towards both types of technologies. However, since constructs' mean scores were quite positive (seven out of eight constructs have a mean score of $\geq$ 5.3/7), this scenario seems less likely. In either case, AI does not stand out among technologies.

6.2. Inconclusive results and proposed improvements

The major reason behind this study's inconclusive results derives from too many uncontrolled variables in this study, which in turn leads to wide variability between subjects. Even though Levene's Test of Homogeneity rarely shows significant results, we observed wide discrepancies between participants during the experiment due to these uncontrolled variables. For example, the interactions with the product and the outcomes of the user test differ greatly between participants. These differences are, in part, due to Babylon's frequent failure of one of basic conversation principles described by Moore et al. (2017): repair. When users take the wrong path at the beginning, or when they answer very incorrectly a question related to their symptoms, Babylon does not offer an alternative path or a way to correct it. Users who notice that they took a wrong turn at some point often complain about being unable to correct their trajectory in the middle of the interaction and being forced to restart the whole process to rectify.

Another uncontrolled variable concerns information pertaining to the product at hand. During the interview, some participants said that they could not trust the answers given to them because they did not know where the information was coming from and how it was sorted. Conversely, curious participants who clicked on at least one of the possible diagnosis Babylon suggested, expressed during the interview a more positive view of the app. They explained that during the experiment, they did not always understand why they were asked certain questions. Looking at the answers helped them understand the process Babylon went through to provide them the suggestions, which increased their trust in the results. This finding seems to support Sharan and Romano's (2020) recommendation to increase transparency and explain how the AI system works to increase trust in AI. However, not all participants looked extensively at the results, in part because they were not specifically asked to. This leads to yet another different interaction with the product, increasing variance – though not statistically significant – between subjects.

Uncontrolled variables blur our vision so much it becomes impossible to say if there are truly no difference in subjective evaluation arising from the

treatment. We could try to counter the effect of these uncontrolled variables by recruiting participants more selectively, but it would only affect users' characteristics, such as English level. It is but one aspect among many others and far from enough to remedy variance. Another way would be to choose another AI-system that we could more easily control. But then, other problems might arise from that device's own specificities. For example, the new device might use a more jargon-based vocabulary, which in turn would also influence user's subjective evaluation. The most secure way to improve effect size might just be to sharply increase the sample size. This way, at a certain point, these differences will washout.

Nevertheless, we think necessary to be more selective regarding participants' mastery of English. Apart from freeing our data from part of its parasitic noise, and thusly require fewer participants, this would also improve data's internal consistency. We suspect low English level bears major responsibility of less-than-ideal Cronbach's alpha. Furthermore, explicitly instructing participants not to take a covid-check test might also improve controllability. Finally, improving instruction quality might also help – as long as they remain short, understandable and rememberable.

6.3. Experimental validity and limitations

Critical look back on this project

From the conception of the experiment, we tried to incorporate counter measures to the threats to User Experience Evaluations (UXE) validity Kieffer (2017) identified. Table 28 presents an overview of the threats our study is susceptible to, our perceived success in counteracting them and a justification.

Statistical conclusion validity		
Statistical power	-	Effect size was mostly small
Fishing and error rate problem	+/-	Test of Normal Distribution and Levene's test for Equality of Variance were consistently conducted. When results were statistically significant, we used nonparametric tests
Reliability of measures	+/-	Use of standardized tests (UEQ+) Cronbach's alpha mostly sufficient, but not always Participants' English level was sometimes barely sufficient to distinguish nuances between items

Reliability of treatment implementation	+	Use of same prepared oral instructions Use of same survey
Random irrelevancies in setting	-	No setting control (participant's home)
Heterogeneity of population	+	Same age group (20-30) Similar education level
Internal validity		
Maturation	+/-	Overall, session time was between 20 and 45 min., but sometimes it was longer
Testing	+	Random assignment to treatment groups
Instrumentation	+	Automated data collection; Same instruments/observers
Statistical regression	+	Random assignment to treatment groups
Selection	+	Random assignment to treatment groups
Mortality	+	One of the groups was missing 1 person due to soaring dropout rates in the second part of the experiment. This was mitigated by randomly assigning participants to both groups
Construct validity		
Poor construct definition	-	UEQ+ constructs are validated but attitude towards technology might not be
Mono-method bias	+	Multiple methods used (UEQ+, interview, attitude test)
Interactions	+	Planning with buffer time between experiments
Construct confounding	+	Several related constructs were studied
Social threats	+	Social interactions were mostly kept to a minimum
External validity		
People	+	Participants were selected randomly and not based on pre-test scores
Setting	+	Experiment replicated in different settings
Time	+	Experiment took place at five different times during the day
Ecological validity		
Noise	-	No control over noise. Sometimes participants were interrupted by roommates or there was interference by outside noise such as construction sounds
Testing location	-	Synchronic remote meeting (participant and experimenter were in their respective homes)
Internet stability	-	No control over internet stability. Sometimes participants had problems with their internet connection, which disrupted their interactions with the experimenter and with the product.
Items (objects)	+	Real objects

Visual refinement	+	Finished product
Dynamicity	+	Implemented
Device	+	Computer
Interaction	+	Performed for real
Action fidelity	-	Imagined scenario
Note : A – symbol signals a possible weakness. A + symbol indicates it is controlled. A +/- symbol means it was mostly controlled with a few exceptions		

Table 28 Threats to the validity (Kieffer, 2017) and our response to counter them

In general, the experiment decently addresses most threats. Threats to internal, external and construct validity are mostly controlled. Problems that we encountered regarding statistical conclusion validity were mostly related to effect size. Indeed, we often found the effect size to range from small to medium as well as a non-normal distribution. Even though we acted on these findings by using appropriate nonparametric tests, the strength of the results is nonetheless diminished if they were to be compared with similar results deriving from a large effect size and normally distributed data. We qualify this as the first major weakness of our experiment.

Even though we tried to recruit as many participants as possible, we did not manage to gather enough data to present clear and undisputable results. This weakness can be attributed to circumstance (covid-19 pandemic). In many ways, it made recruitment harder since it could only take place through online means. The inherent restrains of a thesis work, meant a lack of time and funding, which also did not help raise participation numbers. The choice of design of our experiment also worked against us: as experimental designs require more participants because all groups need to be normally distributed and need to have a medium to large effect size. When compared to a single-group experiment, the chosen experimental design requires at least double the participants.

The second main weakness of this experiment derives from poor control of the setting, which ensues poor ecological validity. By being forced to conduct our experiment remotely, we were unable to provide our participants with a fully standardized environment. Some participants were interrupted by roommates or family members during the experiment, while others had to deal with external noise, such as noise from construction sites.

Furthermore, participants sometimes had internet connection related problems, which increased the duration of the survey, buffer time between screens (during the use of the product) and the time it took to test the product overall. Sometimes, this also impacted the interview – participants and the experimenter were, at times, cut off from each other. Even though these events only meddled with a small sample of our population, when it occurred, the lack of control over the setting could potentially have a far-reaching impact on the results, in unmeasurable ways, and thusly represents a significant flaw in the execution of the experiment.

Additionally, the 'action fidelity' aspect was not thoroughly respected. Participants were often unable to remember the symptoms they had to input into Babylon and as a direct result of that, they were frequently switching between the Babylon app and the survey, where they could find the description of the situation. This back-and-forth struggle coupled with the difficulty and weirdness associated with using a health app in a foreign language, more than likely directly influenced participant's experience of the product. Furthermore, this experiment was conducted during the second COVID-19 lockdown in Belgium. During this time health-related issues were placed in the spotlight like never before, and so, given this particular context, the healthcare nature of the app might have influenced participant's views of the product in ways it would not have under 'normal' circumstances. Additionally, participants who were a part of the treatment groups were reminded fourteen times (twice explicitly and twelve times implicitly) that they were interacting with an AI-enabled system. This constant, albeit passive reminder, might also have induced an unintended change in behaviour and perception that would not have occurred during a normal use of the product.

Another shortcoming of the study relates to the lack of framework validity surrounding studies on UX of medical AI. Firstly, we found no agreed-upon scale to efficiently measure attitude towards technology. These types of scales should be incorporated into UX studies regarding AI, as they would provide important insights regarding the participants. Secondly, as Campbell (2020) noted, UX-AI lacks clear frameworks, stemming in part, from different understandings of the same terms. To avoid using multiple

frameworks that would lead to noncomparable results across studies, we used Schrepp and Thomaschewski (2019) standardized construct definition and testing method. Yet, our research question having not yet been a part of a similar study, to the extent of our knowledge, inevitably means that we have no way to compare and contrast our methods and results. Hopefully, future research will be able to dig further and find new and fruitful insights.

Recommendations moving forward

Experimental design has proven in the past to be an effective format to uncover differences between groups. The absence of significant results in this study is not caused by the inadequacy of the experiment's design. Detecting the differences in behaviour, perception and treatment remains an important subject to explore, since this information enables UX practicians to design counter strategies tailored for medical AI apps' unique needs. Furthermore, the combination of scale questions (either on a computer or on paper) such as the UEQ+ and short interviews should be continued. Scale questions allow fast, easily comparable and quantifiable data, which is especially useful when calculating significant statistical differences, when comparing results with other studies or when attempting to get the full picture. Short interviews allow for personal expression, repair (if needed) when something is not properly understood, new leads, which ultimately induce meaningful insights not provided by close-ended questions. Combining the two enhances the quality of the results since they stem from the best of both worlds. Doing so is, however, considerably time-consuming and financially draining.

Ideally, to obtain true action fidelity, similar research should be conducted on users who are genuinely sick at the time of the experiment and therefore would not feel the need to imagine symptoms but would have to rather simply relay them. Using the product would then likely seem easier and more natural, because they would not have to follow a scenario thusly decreasing the error rate related to symptom input. Most importantly, all reservations regarding effects such as users' uniqueness neglect concerns, if applicable, would then manifest themselves. Such a study would require of researchers to either travel to doctors' practices or hospital grounds.

To obtain actionable results, augment effect size and reach normal distribution, such study requires a much larger pool of participants. We estimate at least 100 participants per treatment group should be recruited to acquire a meaningful size effect with a Confidence Interval of 95%.

Lastly, research should be conducted in a lab to increase controllability of environmental variables. Even though the lab setting might affect users' evaluation, users' normal setting makes them prone to having a bad internet connection, interruptions during the experiment by roommates or family members, different types of malfunctions interfering with the interactions, which in turn also affects users' evaluation. A controlled environment may mimic users' normal setting to an extent and gather mostly uncompromised data, but an uncontrolled environment can completely corrupt the data and destroy any hope of replicability.

7. CONCLUSION

Results from this study are mostly inconclusive but still manage to contribute to research in two important ways. Firstly, Gustafsod's theory on gender biases as pertaining to risk assessment might need to be revisited. Differences between genders may be eroding and are possibly becoming less relevant and appropriate to use when dealing with younger generations. They might also be much more context-dependent than previously thought: a hot topic such as health during a pandemic might soften or override gender-based variation in perception to such an extent it might even become irrelevant. Secondly, differences in subjective evaluation between AI-related products and non-AI related products are not inherently negative towards AI-related products. Some aspects, like the construct novelty, might in fact benefit from the excitement that comes with possibly ground-breaking technologies. Further research on the positive influence of AI-labelling might open up new paths and techniques which might be useful if we are to tackle low adoption rates.

The interviews we conducted for this study revealed that trust is indeed at the forefront of users' assessment of apps. When users do understand the 'motives' behind AI's proceedings, they trust the results more. Furthermore, a lack of trust in the source of the data which is being used by the AI or in

the company producing the AI or the data also plummets users' propensity to adopt the product, regardless of the AI's capacities. When attempting to increase adoption rates, users' trust is a *must have*. Without trust in the mHealth app, users outright refuse to use it and instead prefer to consult a medical doctor.

That is not to say that all these reluctances to adopt AI systems require 'fixing' – since, not all of them are inherently bad. Users with critical views on data quality, data origin and application owners should be celebrated, because such views make for informed decisions. Users making decisions based on wrong information, obtained from a poorly trained AI system or from private firms whose sole desire is to profit from their medical condition. On top of having the potential for undesirable or sometimes even devastating consequences for the lives of these users, this would worsen the healthcare crisis even further. Being aware of these reluctances enables UX designers to target users' concerns and respond to them both appropriately and ethically by, for example, mentioning partnership with local hospitals and practitioners.

Much more useful and practical information can be gleaned from future such studies. Moreover, more data should be harvested and thoroughly analysed to fully facilitate the shift towards Smart Healthcare. Even though this thesis showed promising design to evaluate differences in subjective evaluation, the context in which it was carried out impacted the quality of the data. Conducting the experiment during the second lockdown of a global pandemic damaged the data that was gathered in two ways. First, the decrease in validity of the data. The ecological validity in particular suffered immensely from the decrease in controllability. Second, the reliability of the results. The pandemic and its health focused information war coupled with our product's focus on health might have influenced users' evaluation due the pandemic circumstances. Anxiety in relation to health being higher than usual, users might evaluate products differently. Apart from the sensitiveness of the subject itself, being unable to move easily and safely in the outside world, as well as having trouble booking appointments with medical professionals, heightens these apps' usefulness in new ways. These changes might be limited in time and pass once the pandemic is officially

over – in which case, it will be back to business as usual, and we will be faced with the same problems. The chance exists though, these changes might also be long term and completely transform users' relation to healthcare. Either way, much more research is required to achieve ethical, sustainable and smart healthcare.

Finally, since healthcare systems require a momentous reform to properly undertake current and future challenges such as an increase in age-related illnesses due to ageing populations and pandemics, UX research must prioritize or at least intensify the focus on AI adoption to face these societal challenges head on. UX research would particularly benefit from clear, flexible and extensive frameworks that can be used in a plethora of contexts to evaluate users' subjective perceptions. Most importantly, researchers would need to rally behind this framework. Gaining insights from cross-study comparisons is quintessential to properly interpret defining components that influence users' subjective evaluation. But, to conceive the framework on which future UX researchers can rely on, the groundwork that will be laid by a study or by different studies would require many more participants than most UX studies usually rely on. This is because acquiring sufficient reliable and complete data requires designs and tests which are very demanding in numbers in order to be truly meaningful. Only after this work is done can UX researchers propose tailored strategies valid for the myriad of AI applications that exist. They would then, when it is all said and done, contribute in unfathomable ways to a more sustainable future.

8. REFERENCES

Abd-Alrazaq, A., Safi, Z., Alajlani, M., Warren, J., Househ, M., & Denecke, K. (2020). Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review. *Journal of Medical Internet Research*, *22*(6), e18301. https://doi.org/10.2196/18301

Alexandre, L. (2017). La mort des médecins. *Les Tribunes de la santé*, *54*(1), 43. https://doi.org/10.3917/seve.054.0043

Alilyyani, B., Wong, C. A., & Cummings, G. (2018). Antecedents, mediators, and outcomes of authentic leadership in healthcare: A systematic review. *International Journal of Nursing Studies*, *83*, 34–64. https://doi.org/10.1016/j.ijnurstu.2018.04.001

Altindis, S. (2011). Job motivation and organizational commitment among the health professionals: A questionnaire survey. *African Journal of Business Management*, *5*(21), 8601–8609. https://doi.org/10.5897/AJBM11.1086

Antonelli, W. (2020, November 18). What is Zoom? A comprehensive guide to the wildly popular video-chatting service for computers and smartphones. *Business Insider*. https://www.businessinsider.com/what-is-zoom-guide?r=US&IR=T

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Barreiro, C. A., & Treglown, L. (2020). What makes an engaged employee? A facet-level approach to trait emotional intelligence as a predictor of employee engagement. *Personality and Individual Differences*, *159*, 109892. https://doi.org/10.1016/j.paid.2020.109892

Bass, B. & Avolio, B. (1992). Multifactor Leadership Questionnaire—Short Form 6S. *Center for Leadership Studies*. Binghamton: New York.

Benton, D. C. (2015). Mapping and Changing Informal Nurse Leadership Communication Pathways in a Health System. *Asian Nursing Research*, *9*(1), 28–34. https://doi.org/10.1016/j.anr.2014.10.006

Buil, I., Martínez, E., & Matute, J. (2019). Transformational leadership and employee performance: The role of identification, engagement and

proactive personality. *International Journal of Hospitality Management*, *77*, 64–75. https://doi.org/10.1016/j.ijhm.2018.06.014

Bulger, M. E., Mayer, R. E., & Metzger, M. J. (2014). Knowledge and processes that predict proficiency in digital literacy. *Reading and Writing*, *27*(9), 1567–1583. https://doi.org/10.1007/s11145-014-9507-2

Cameron, G., Cameron, D. W., Megaw, G., Bond, R. R., Mulvenna, M., O'Neill, S. B., Armour, C., & McTear, M. (2018, July 1). *Best Practices for Designing Chatbots in Mental Healthcare – A Case Study on iHelpr*. Proceedings of the 32nd International BCS Human Computer Interaction Conference. https://doi.org/10.14236/ewic/HCI2018.129

Campbell, J. L. (2020). Healthcare experience design: A conceptual and methodological framework for understanding the effects of usability on the access, delivery, and receipt of healthcare. *Knowledge Management & E-Learning*, *12*(4), 505–520. https://doi.org/10.34105/j.kmel.2020.12.028

Campbell, D. T., & Stanley, J. C. (1967). *Experimental and quasi-experimental designs for research* (2. print; Reprinted from "Handbook of research on teaching"). Houghton Mifflin Comp.

Chowriappa, P., Dua, S., & Todorov, Y. (2014). Introduction to Machine Learning in Healthcare Informatics. In S. Dua, U. R. Acharya, & P. Dua (Eds.), Machine Learning in Healthcare Informatics (pp. 1–23). Springer. https://doi.org/10.1007/978-3-642-40017-9_1

Chung, K., & Park, R. C. (2019). Chatbot-based heathcare service with a knowledge base for cloud computing. *Cluster Computing*, *22*(S1), 1925–1937. https://doi.org/10.1007/s10586-018-2334-5

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). New Jersey: Lawrence Erlbaum.

Correia, F., Alves-Oliveira, P., Maia, N., Ribeiro, T., Petisca, S., Melo, F. S., & Paiva, A. (2016). Just follow the suit! Trust in human-robot interactions during card game playing. *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 507–512. https://doi.org/10.1109/ROMAN.2016.7745165

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of*

Marketing Science, *48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

de Gennaro, M., Krumhuber, E. G., & Lucas, G. (2020). Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood. *Frontiers in Psychology*, 10, 3061. https://doi.org/10.3389/fpsyg.2019.03061

Denecke, K., & Warren, J. (2020). How to Evaluate Health Applications with Conversational User Interface? *Studies in Health Technology and Informatics*, 270, 976–980. https://doi.org/10.3233/SHTI200307

Derèze, G. (2019). Méthodes empiriques de recherche en information et communication (2nd ed.). De Boeck Supérieur.

Distler, V., Lallemand, C., & Bellet, T. (2018). Acceptability and Acceptance of Autonomous Mobility on Demand: The Impact of an Immersive Experience. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–10. https://doi.org/10.1145/3173574.3174186

Dorsey, E. R., & Topol, E. J. (2020). Telemedicine 2020 and the next decade. *The Lancet*, *395*(10227), 859. https://doi.org/10.1016/S0140-6736(20)30424-4

Dua, S., Acharya, U. R., & Dua, P. (Eds.). (2014). *Machine Learning in Healthcare Informatics* (Vol. 56). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-40017-9

Ejdys, J. (2018). Trust in Technology in Case of Humanoids Used for the Care for the Senior Persons. *Multidisciplinary Aspects of Production Engineering*, *1*(1), 875–881. https://doi.org/10.2478/mape-2018-0110

Fan, W., Liu, J., Zhu, S., & Pardalos, P. M. (2020). Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annals of Operations Research*, *294*(1–2), 567–592. https://doi.org/10.1007/s10479-018-2818-y

Fraser, H., Coiera, E., & Wong, D. (2018). Safety of patient-facing digital symptom checkers. *The Lancet*, *392*(10161), 2263–2264. https://doi.org/10.1016/s0140-6736(18)32819-8

George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Boston: Allyn & Bacon.

Gille, F., Jobin, A., & Ienca, M. (2020). What we talk about when we talk about trust: Theory of trust for AI in healthcare. *Intelligence-Based Medicine*, *1–2*, 100001. https://doi.org/10.1016/j.ibmed.2020.100001

Giraud, L., Autissier, D., Johnson, K. J., & Moutot, J.-M. (2013). Attitudes et comportements des salariés envers le changement: Une étude longitudinale de la mise en place d'un changement organisationnel. *Question(s) de management*, *3*(2), 37. https://doi.org/10.3917/qdm.132.0037

Grice, P. (1975). Logic and conversation. In Cole, P.; Morgan, J. (eds.). Syntax and semantics. 3: Speech acts. New York: Academic Press. pp. 41–58.

Gruson, D., Helleputte, T., Rousseau, P., & Gruson, D. (2019). Data science, artificial intelligence, and machine learning: Opportunities for laboratory medicine and the value of positive regulation. *Clinical Biochemistry*, *69*, 1–7. https://doi.org/10.1016/j.clinbiochem.2019.04.013

Gustafsod, P. E. (1998). Gender Differences in risk perception: Theoretical and methodological perspectives. *Risk Analysis*, *18*(6), 805–811. https://doi.org/10.1111/j.1539-6924.1998.tb01123.x

Hassenzahl, M. (2003). The thing and I: Understanding the relationship between user and product. In M. A. Blythe, K. Overbeeke, A. F. Monk, & P. C. Wright (Eds.), *Funology* (Vol. 3, pp. 31–42). Springer Netherlands. https://doi.org/10.1007/1-4020-2967-5_4

Heudel, P.-É., Durand, T., & Blay, J.-Y. (2017). Projets d'intelligence artificielle à l'échelle d'un établissement de santé: L'exemple du centre Léon Bérard. *Revue française des affaires sociales*, *1*(4), 133. https://doi.org/10.3917/rfas.174.0133

Higgins, D., & Madai, V. I. (2020). From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. *Advanced Intelligent Systems*, *2*(10), 2000052. https://doi.org/10.1002/aisy.202000052

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407-434. https://doi.org/10.1177/0018720814547570

Hoffmeyer-Zlotnik, J. (2016). Harmonisation of Demographic and Socio-Economic Variables in Cross-National Survey Research. *GESIS Survey Guidelines*. Mannheim, Germany: GESIS -Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_012

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & McTear, M. (2019). Usability testing of a Healthcare Chatbot: Can we use conventional methods to assess conversational User Interfaces? *Proceedings of the 31st European Conference on Cognitive Ergonomics*, 207–214. https://doi.org/10.1145/3335082.3335094

Johnson, H. (2020). 10 Best AI Based Healthcare Apps You Can Try in 2020. *SwissCognitive - The Global AI Hub*. Retrieved January 11, 2021, from https://swisscognitive.ch/2020/03/27/10-best-ai-based-healthcare-apps-you-can-try-in-2020/

Kieffer, S. (2017). ECOVAL: Ecological Validity of Cues and Representative Design in User Experience Evaluations. *AIS Transactions on Human-Computer Interaction*, *9*(2), 149-172. https://doi.org/10.17705/1thci.00093

Komi, A. K. (2019). Le management des résistances à un projet d'innovation par l'intelligence artificielle dans une perspective de changement. *RIMHE: Revue Interdisciplinaire Management, Homme & Entreprise*, *36*(3), 29. https://doi.org/10.3917/rimhe.036.0029

Lallemand, C. & Koenig, V. (2017). "How Could an Intranet be Like a Friend to Me?" – Why Standardized UX Scales Don't Always Fit. *Proceedings of ECCE 2017*, Umeå, Sweden

Law, E. L.-C., van Schaik, P., & Roto, V. (2014). Attitudes Towards User Experience (UX) Measurement. *International Journal of Human-Computer Studies*, *72*(6), 526–541. https://doi.org/10.1016/j.ijhcs.2013.09.006

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lew, G. & Schumacher, R. M. (2020). *AI and UX: Why Artificial Intelligence Needs User Experience.* Apress. https://doi.org/10.1007/978-1-4842-5775-3

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Mishra, P., Pandey, C., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive Statistics and Normality Tests for Statistical Data. *Annals of Cardiac Anaesthesia*, *22*(1), 67–72. https://doi.org/10.4103/aca.ACA_157_18

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital Health*, *5*. https://doi.org/10.1177/2055207619871808

Norman, D. A. (2005). *Emotional design: Why we love (Or hate) everyday things*. Basic Books.

Peters, C. A. (2001). Statistics for Analysis of Experimental Data. In Powers, S. E. (Ed.), *Environmental Engineering Processes Laboratory Manual*. AEESP, Champaign, IL.

O'Hear, S. (2017, January 4). Babylon Health Partners with UK' NHS to Replace Telephone Helpline with AI-Powered Chatbot. TechCrunch. Retrieved March 15, 2021, from https://social.techcrunch.com/2017/01/04/babylon-health-partners-with-uks-nhs-to-replace-telephone-helpline-with-ai-powered-chatbot/

Ongena, Y. P., Haan, M., Yakar, D., & Kwee, T. C. (2020). Patients' views on the implementation of artificial intelligence in radiology: Development and validation of a standardized questionnaire. *European Radiology*, *30*(2), 1033–1040. https://doi.org/10.1007/s00330-019-06486-0

Panesar, A. (2019). *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes* (1st ed. edition). Apress.

Petrova, V. (2020). *Can we trust our health in the hands of chatbots?: An exploratory study investigating the effect of anthropomorphic design of e-Health chatbots on patients UX.* (public) [Bachelor thesis, University of Twente, Netherlands]. http://essay.utwente.nl/81683/

Pettigrew, A. & Whipp, R. (1993). Managing Change for Competitive Success. Blackwell Publishing. ISBN: 978-0-631-19142-1

Rauschenberger, M., Schrepp, M., Perez-Cota, M., Olschner, S., & Thomaschewski, J. (2013). Efficient Measurement of the User Experience of Interactive Products. How to use the User Experience Questionnaire (UEQ). Example: Spanish Language Version. *International Journal of Interactive Multimedia and Artificial Intelligence*, *2*(1), 39. https://doi.org/10.9781/ijimai.2013.215

Rourke, L., & Anderson, T. (2004). Validity in Quantitative Content Analysis. *Educational Technology Research and Development*, *52*(1), 5-18. Retrieved May 21, 2021, from http://www.jstor.org/stable/30220371

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). The Measurement of Work Engagement with a Short Questionnaire. *Educational and Psychological Measurement*, *66*(4), 701-716 https://doi.org|10.1177/0013164405282471

Schrepp, M. & Thomaschewski, J. (2019b). *UEQ+: A Modular Extension of the User Experience Questionnaire*. UEQ plus. Retrieved February 3rd, 2021, from https://ueqplus.ueq-research.org/

Schrepp, M., & Thomaschewski, J. (2019a). Design and Validation of a Framework for the Creation of User Experience Questionnaires. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(7), 88. https://doi.org/10.9781/ijimai.2019.06.006

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(6), 103. https://doi.org/10.9781/ijimai.2017.09.001

Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, *6*(8), e04572. https://doi.org/10.1016/j.heliyon.2020.e04572

Sheikh, A., Sood, H. S., & Bates, D. W. (2015). Leveraging health information technology to achieve the "triple aim" of healthcare reform. *Journal of the American Medical Informatics Association*, *22*(4), 849–856. https://doi.org/10.1093/jamia/ocv022

Solon, O. (2014, April 28). Babylon app puts a GP in your pocket. *WIRED UK*. Retrieved January 16, 2021, from https://www.wired.co.uk/article/babylon-ali-parsa

Thüring, M., & Mahlke, S. (2007). Usability, aesthetics and emotions in human–technology interaction. *International Journal of Psychology*, *42*(4), 253–264. https://doi.org/10.1080/00207590701396674

Tien, J. M. (2017). Internet of Things, Real-Time Decision Making, and Artificial Intelligence. *Annals of Data Science*, *4*(2), 149–178. https://doi.org/10.1007/s40745-017-0112-5

Turner, P., Kushniruk, A., & Nohr, C. (2017). Are We There Yet? Human Factors Knowledge and Health Information Technology – the Challenges of Implementation and Impact. *Yearbook of Medical Informatics*, *26*(01), 84–91. https://doi.org/10.15265/IY-2017-014

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). 'Do you trust me?': Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 7–9. https://doi.org/10.1145/3308532.3329441

White, B. A. (2014, April 28). Tech pioneer brings GP appointments into the living room. The Telegraph. Retrieved January 16, 2021, from https://www.telegraph.co.uk/finance/businessclub/10793175/Tech-pioneer-brings-GP-appointments-into-the-living-room.html

Yang, Q., Steinfeld, A., & Zimmerman, J. (2019). Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3290605.3300468

Yue, C. A., Men, L. R., & Ferguson, M. A. (2019). Bridging transformational leadership, transparent communication, and employee openness to change: The mediating role of trust. *Public Relations Review*, *45*(3), 101779. https://doi.org/10.1016/j.pubrev.2019.04.012

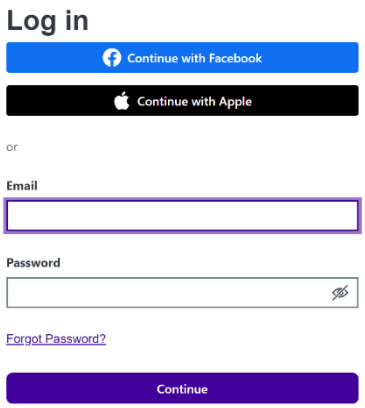
Zeitoun, J.-D., & Ravaud, P. (2019). L'intelligence artificielle et le métier de médecin. *Les Tribunes de la santé*, *N° 60*(2), 31. https://doi.org/10.3917/seve1.060.0031

9. APPENDICES

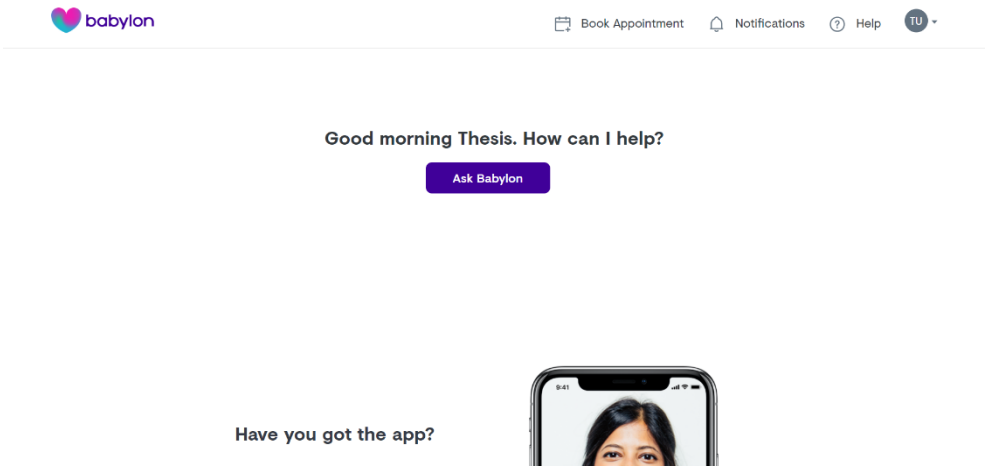
9.1. Babylon presentation

In the following section, we present Babylon's interface through screenshots. We show the main screens participants saw during the experiment.

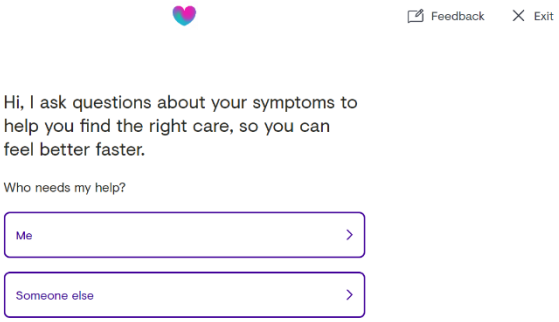
9.1.1. Log-in



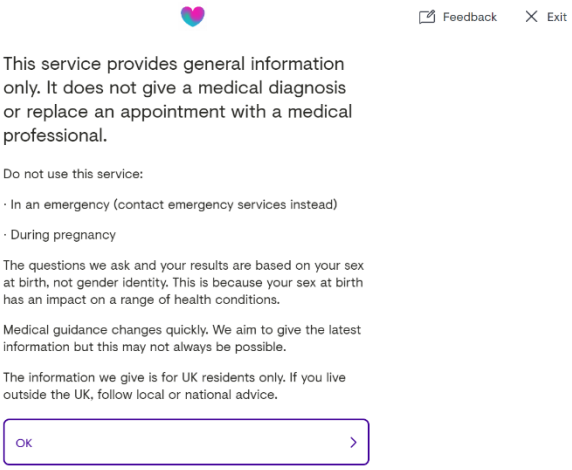
9.1.2. Home page



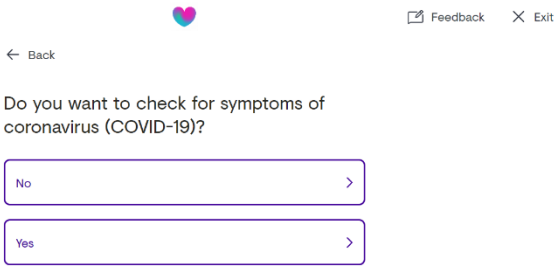
9.1.3. First question



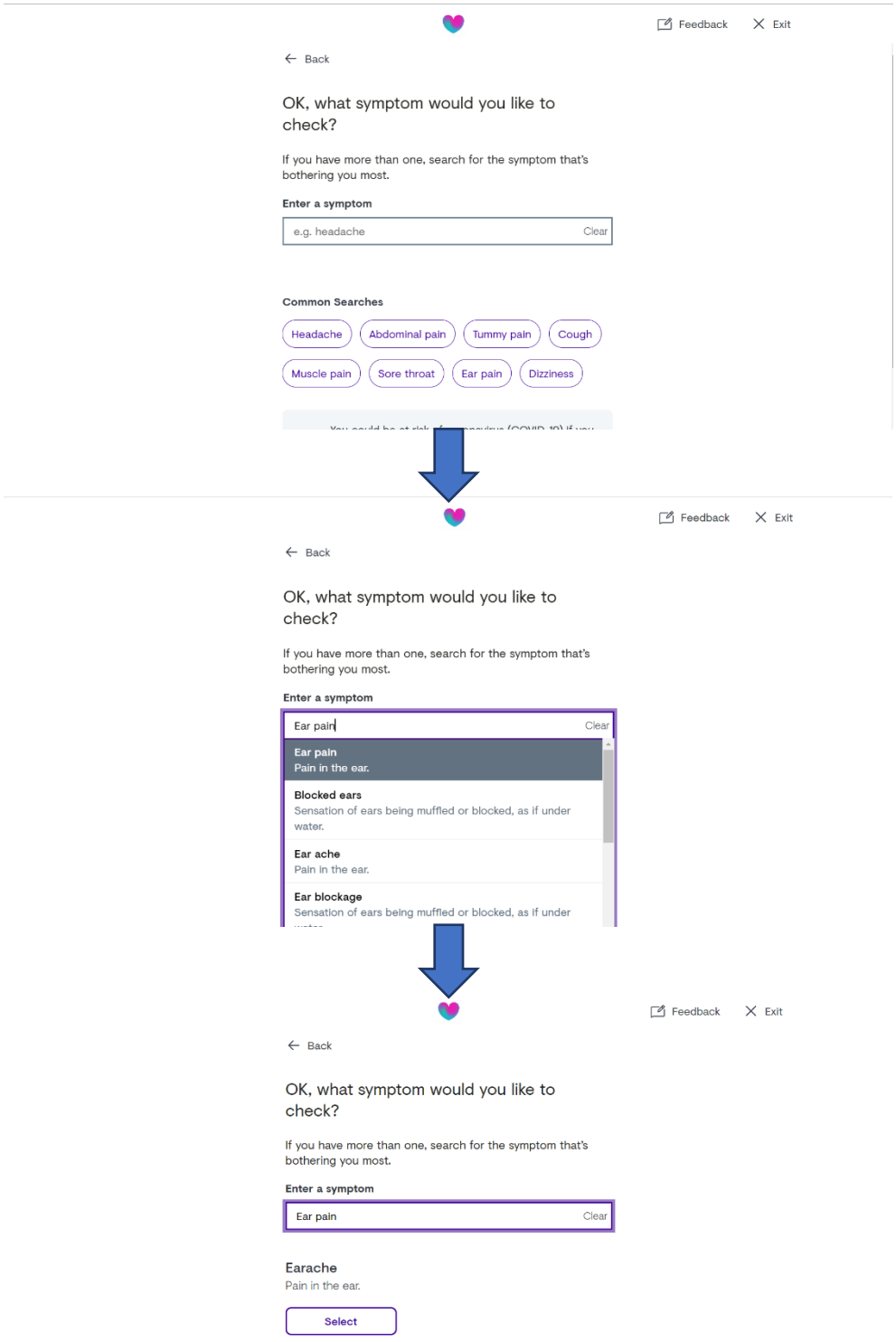
9.1.4. Disclaimer



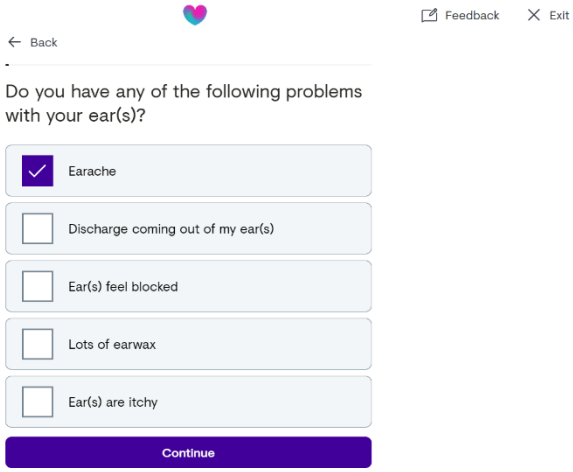
9.1.5. Covid related question



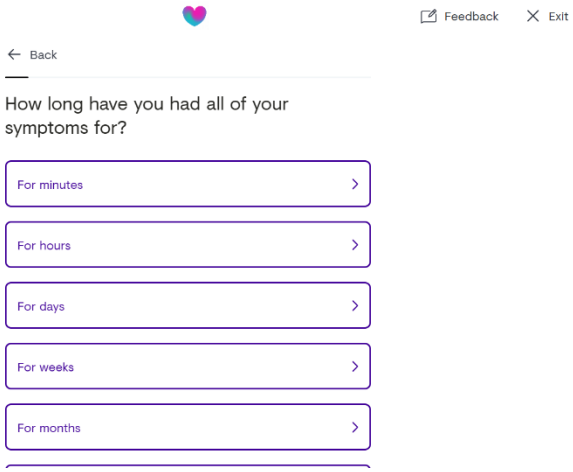
9.1.6. Entering the first symptom (three steps)



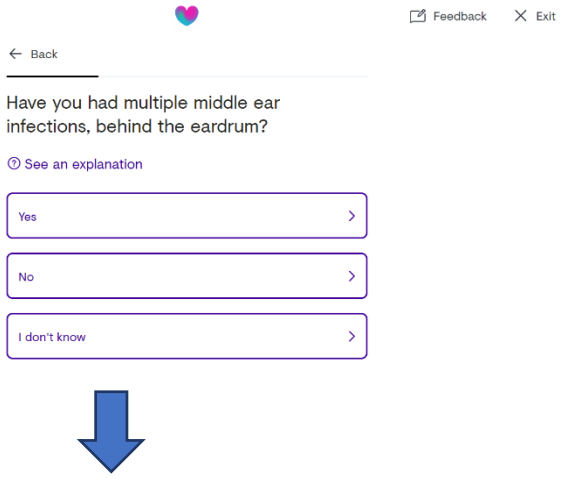
9.1.7. Question format 1

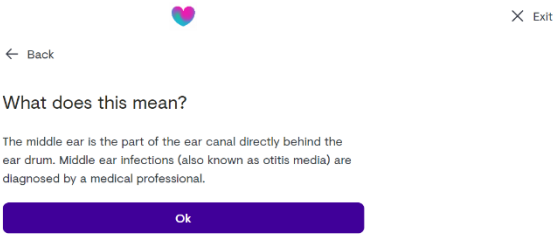


9.1.8. Question format 2

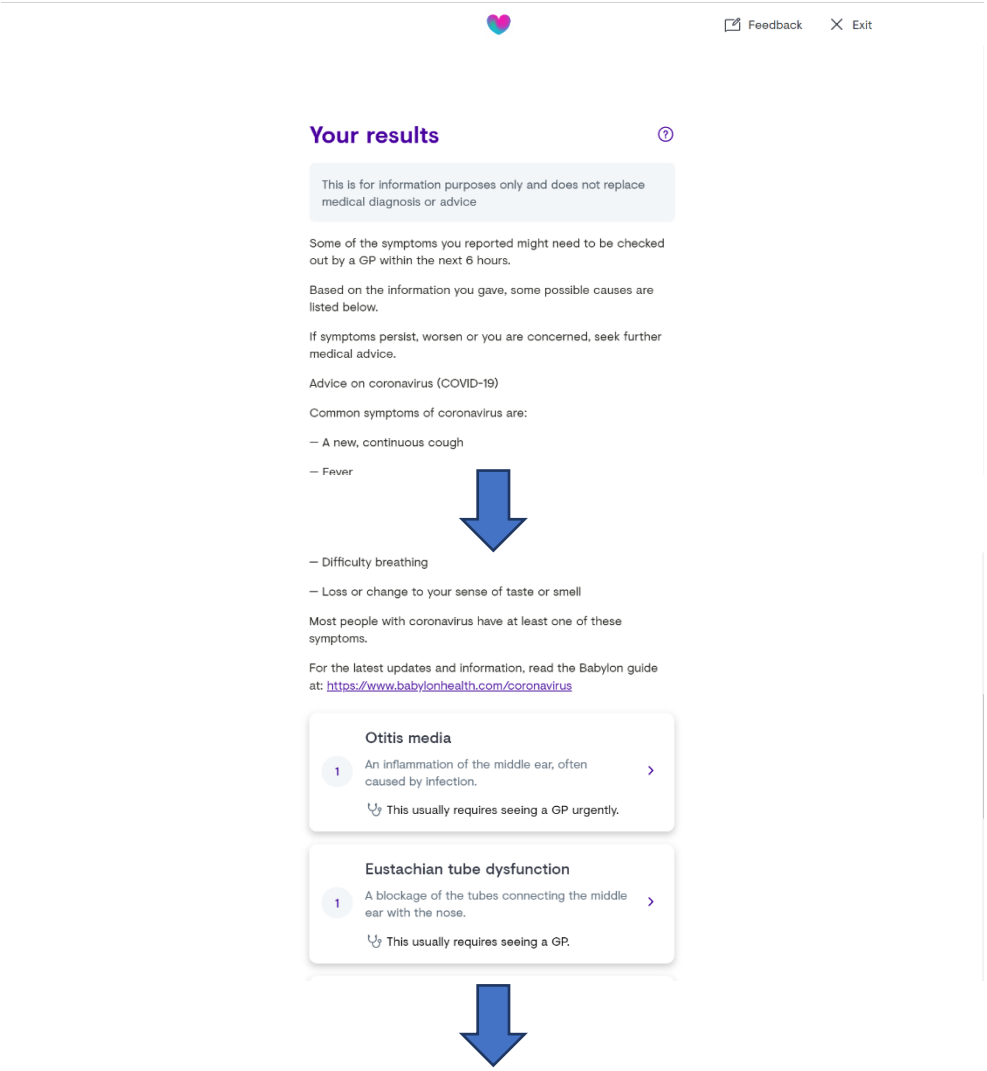


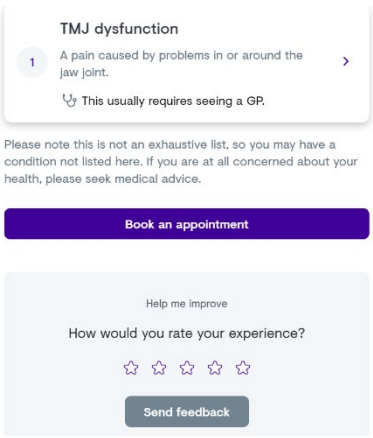
9.1.9. "See an explanation" button and button clicked



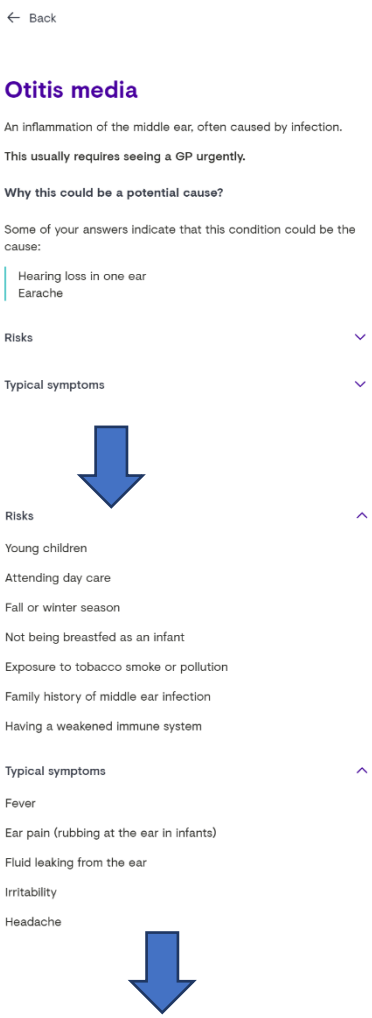


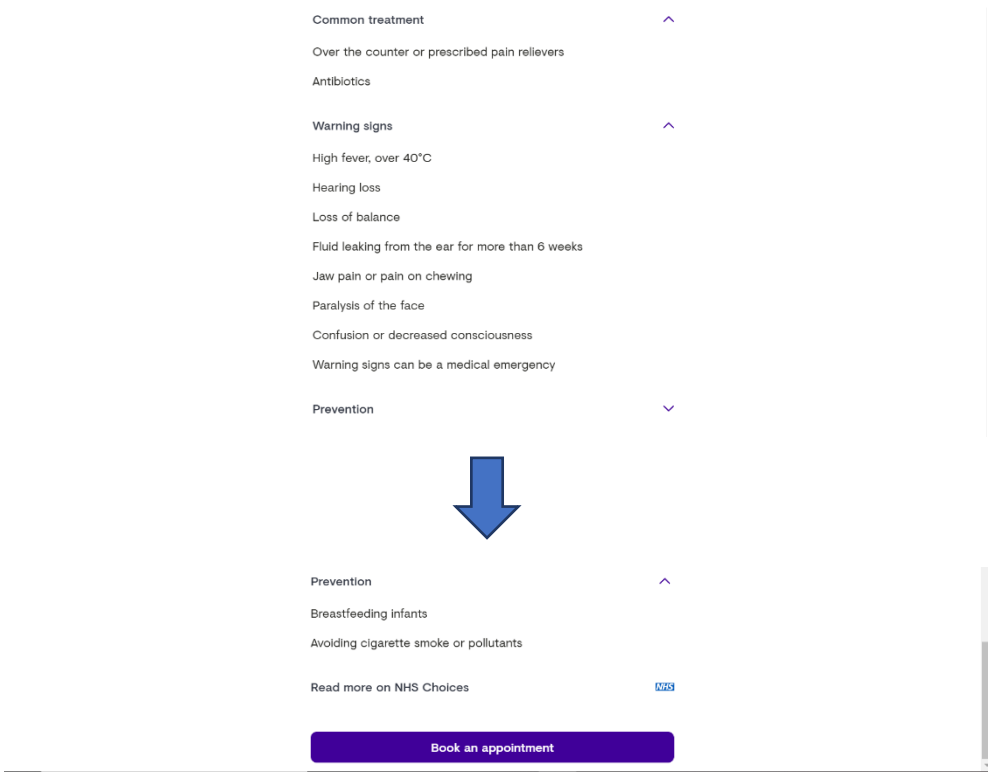
9.1.10. Result page (simulated scrolling)





9.1.11. Detailed diagnosis (simulated with scrolled page and clicked headings)




Common treatment
Over the counter or prescribed pain relievers
Antibiotics
Warning signs
High fever, over 40°C
Hearing loss
Loss of balance
Fluid leaking from the ear for more than 6 weeks
Jaw pain or pain on chewing
Paralysis of the face
Confusion or decreased consciousness
Warning signs can be a medical emergency
Prevention

Prevention
Breastfeeding infants
Avoiding cigarette smoke or pollutants
Read more on NHS Choices
Book an appointment

9.2. Participant recruitment

9.2.1. Consent form for LCOMU2812 participants

Informed consent

Participation in a master's thesis in UX

1. **Introduction and Purpose of the Study**

You are being invited to participate in a research study about user experience of an online product. This research project is being conducted by final-year student Daniela Magalhaes Azevedo and supervised by Pr. Dr. Suzanne Kieffer (UCLouvain). The objective of this research project is to attempt to understand how and why people experience this product differently.

2. **Description of the Research**

If you agree to participate in this research study, the following will occur: at the start of the experiment, you are given a description of a specific situation. This is the setting in which you are asked to use the product. With this situation in mind, you then proceed to use the product. Once you finish using the product, we ask you to evaluate your experience. More specifically, we ask you to recall your interaction with the product and share your opinion. Then, we ask you to give us your general impression on subjects related to the product and a few details about yourself.

During the experiment, we ask you to turn on your camera and share your screen with us.

3. **Subject Participation**

We estimate that 25 participants will enrol in this study. Participation will involve one online session, approximately 30 minutes in length.

During this experiment, participants must have access to a computer, a stable internet connection and their webcam switched on.

An upper intermediate English level (B2) is required to successfully complete this experiment.

4. **Potential Benefits**

Participation in this study will be considered when grading your LCOMU2812 course.

5. **Potential Risks and Discomforts**

There are no known risks.

6. **Questions**

If you have any questions about the study, please contact Daniela Azevedo by email at daniela.magalhaes@student.uclouvain.be.

7. **Confidentiality**

All information taken from the study will be coded to protect each participant's name. No names or other identifying information will be used when discussing or reporting data. The researcher will safely keep all files and data collected in an encrypted file on

her personal electronic device. Once the data has been fully analysed, it will be destroyed.

8. Consent

Your decision to participate in this study is completely voluntary. If you decide to participate in this study, you can withdraw your consent and discontinue participation at any time without prejudice.

Please read the following statement and click on your preferred choice.

I have carefully read the information that was presented to me and I voluntarily agree to participate in this study. This commitment can be signed and delivered by electronical transfer or by a nominative electronic agreement given on Moodle. I understand that this e-signature has the same value as my written signature.

- ☐ I agree to participate in this study.
- ☐ I do not agree to participate in this study.

9.2.2. Consent form for other participants

Informed consent

Participation in a master's thesis in UX

Introduction and Purpose of the Study

You are being invited to participate in a research study about the User Experience of an online product. This research project is being conducted by final-year student Daniela Magalhaes Azevedo and supervised by Pr. Dr. Suzanne Kieffer (UCLouvain). The objective of this research project is to understand how and why people experience this product differently.

Description of the Research

If you agree to participate in this research study, the following will occur: at the start of the experiment, you are given a description of a specific situation. This is the setting in which you are asked to use the product. With this situation in mind, you then proceed to use the product. Once you finish using the product, we ask you to evaluate your experience. More specifically, we ask you to recall your interaction with the product and share your opinion. Then, we ask you to give us your general impression on subjects related to the product and a few details about yourself.

During the experiment, we ask you to turn on your camera and share your screen with us. The experiment is recorded.

Subject Participation

We estimate that 25 participants will enrol in this study. Participation will involve one online session, approximately 30 minutes in length.

During this experiment, participants must have access to a computer, a stable internet connection and their webcam switched on.

An upper intermediate English level (B2) is required to successfully complete this experiment.

Potential Benefits

Participation in this study will help the scientific community to better understand links between social constructs (i.e., ideas shared by a society) and individual experience.

Potential Risks and Discomforts

There are no known risks.

Questions

If you have any questions about the study, please contact Daniela Azevedo by email at daniela.magalhaes@student.uclouvain.be

UCLouvain

Confidentiality

All information taken from the study will be coded to protect each participant's name. No names or other identifying information will be used when discussing or reporting data. The researcher will safely keep all files and data collected in an encrypted file on her personal electronic device and UCLouvain's encrypted servers. Once the data has been fully analysed, it will be destroyed.

Consent

Your decision to participate in this study is completely voluntary. If you decide to participate in this study, you can withdraw your consent and discontinue participation at any time without prejudice.

Participant's Statement

I, undersigned [redacted], have carefully read the information that was presented to me and I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I understand that my e-signature has the same value as my written signature.

Date:

Participant's signature

Experimenter's signature

UCLouvain

9.2.3. Standard e-mail with instructions

Concerne : expérience mémoire en communication

Bonjour [Prénom],

Merci pour votre intérêt pour cette recherche ! Vous avez dû recevoir un message automatique de Zoom après vous être inscrit.e sur le doodle. Dans ce mail, vous trouverez un lien vers la salle de réunion. Le jour-J, cliquez sur ce lien pour entrer dans la salle et débuter l'entretien. De préférence, installez Zoom au préalable.
Je vous envoie en pièces-jointes un document de consentement éclairé. Ce document de deux pages présente l'expérience plus en détail. Avant le jour-J, il faudra prendre connaissance de ce document, le signer et me le retourner par mail. Vous pouvez le signer électroniquement grâce à un programme comme Adobe, ou vous pouvez l'imprimer, le signer à la main, puis prendre une photo ou le scanner pour ensuite me le renvoyer.

En vous souhaitant une agréable journée,
Daniela Azevedo

9.2.4. Public recruitment message for female participants

Hello les filles !
Dans le cadre de mon mémoire en communication à l'UCLouvain, je cherche des personnes entre 18 et 30 ans pour tester un produit lors d'un entretien en ligne (sur ZOOM) et répondre à quelques questions par rapport à votre expérience et vos ressentis. Je cherche des femmes avec un niveau B2 en anglais, accès à un ordi, un internet stable, une webcam et du son (micro + audio). Au total, l'expérience dure +/- 30minutes.
Si vous êtes motivées, inscrivez-vous à une date qui vous convient sur le doodle. Je vous enverrai ensuite un document à signer et à remettre avant le jour de l'expérience.
https://doodle.com/mm/409443017/ux-thesis-uclouvain

9.2.5. Word-of-mouth recruitment message

Dans le cadre de mon mémoire en communication à l'UCLouvain, je cherche des personnes entre 20 et 30 ans pour tester un produit lors d'un entretien en ligne (sur ZOOM). Après avoir testé le produit, il faudra répondre à un questionnaire en rapport à votre expérience et vos impressions du produit, et à quelques questions sur vos opinions et vos intérêts généraux. Pour participer il faut avoir, un niveau B2 en anglais, accès à un ordi, un internet stable, une webcam et du son (micro + audio). Au total, l'expérience dure +/- 30minutes.

Si vous êtes dispo, inscrivez-vous à une date qui vous convient sur le doodle. Insérez-y une adresse mail valide afin que je vous envoie le document de consentement éclairé. Ce document est à signer et à remettre avant le jour de l'expérience.

https://doodle.com/mm/409443017/ux-thesis-uclouvain

9.3. Pilot experience

9.3.1. Protocol

Given the safety measures in place in early 2021, the experiment takes place on Zoom. During the experiment, we ask participants to turn on their webcam and to share their screen. We record everything to capture facial expressions and impressions, and complement ordinal data collected through the questionnaires.

We randomly assign our participants in to two groups: a No Treatment Group and a Treatment Group. The sole difference between these groups is how we present Babylon Health. In description given to the No Treatment Group, we do not mention artificial intelligence. In the description given to the Treatment Group, we specify that the product is made of artificial intelligence. In Table 29 below, we underline the addition solely for reading purposes – it is not underlined in the document given to the participants.

No Treatment Group	Treatment Group
In this experiment we would like to understand what a user experiences when interacting with a product called Babylon Health. Babylon is a British website where you can enter your symptoms and, based on your input, it gives you possible diagnostics. It doesn't replace a doctor, but it gives you an idea of what you could have, and it advises you on the next steps to take.	In this experiment we would like to understand what a user experiences when interacting with a product called Babylon Health. Babylon is a British website where you can enter your symptoms and, based on your input, its artificial intelligence gives you possible diagnostics. It doesn't replace a doctor, but it gives you an idea of what you could have, and it advises you on the next steps to take.

Table 29 Babylon Health's description for each group

To avoid inducing biases and therefore contaminating our data, we reduce interactions to a bare minimum. Thus, we detail the experiment and give the instructions exclusively on paper. Participants should only contact the experimenter when they encounter an error, they are unsure what is expected of them or when they finish the last questionnaire.

These are the instructions they received in the first page:

To understand the different experiences a user might encounter, we have designed three specific situations. Each situation addresses a particular feature of the product, and thus a different use.

The experiment will go as follows:

First, you will find a description of a situation: this is the setting in which you use the product. The description is full of symptoms relevant to a diagnosis. Imagine you are in that situation. Then, when prompted, please go to the website.

Please read each description of the situation carefully, as it gives you all the relevant information. If something is not mentioned, you can assume it doesn't exist in this situation.

Example:

Babylon asks you if there is a fever. However, the description of the situation doesn't mention anything about temperature. You can therefore assume there is no fever.

You can freely switch back and forth between this document and the website, if you don't remember precisely the information given in the situation. However, you should not use your browser for anything else. Once Babylon has given you your diagnosis and/or suggested next steps to take, please come back to this document and complete the evaluation questionnaire.

If, after reading the instructions, you have a question or something is not working as it should, please tell me immediately by chat or by voice.

The experience is divided in two parts: test and post-test. During the test, participants are tasked to find a diagnosis for three situations using only Babylon Health (Table 30).

N	Description of the situation
1.	It is high season for your seasonal allergies. Your nose has been clogged for the past week, so you have been breathing mostly through your mouth. Your jaw hurts a little and you have been feeling stab-like pain in your right ear since yesterday. Today, you noticed that you were hearing less from that same ear. Your ear feels moist, but nothing has come out. When you look at it, everything seems normal. You have no other symptoms. Your ear bothers you a lot, but you don't want to immediately call the doctor. With only this information in mind, you check your symptoms on the website https://www.babylonhealth.com/ask-babylon-chat.
2.	You have been at home for the past three days. Today, from the moment you woke up, you have been feeling quite tired, even though you went to bed and woke up at the same hour you usually do. As the day goes by, your body increasingly aches all over. As a precaution, you take your temperature every 2 hours. Last time you checked, you had 37.5°C. After taking your contraceptive pill, you notice that your throat starts to feel funny, like there is something stuck that you cannot quite sallow. It aches slightly. You suspect you got the flu because your windows don't fully close, and cold wind passes through. Outside, the temperature has dropped dramatically since yesterday. At the same time, you work a job that requires frequent contact with other people. No one else in your house has similar symptoms. Even though you are not particularly vulnerable to covid-19, because you have no underlying conditions, you are a bit worried, and you want to check your symptoms. You decide to check the website https://www.babylonhealth.com/ask-babylon-chat.
3.	Five days ago, your 25-year-old male friend, Jake, told you he has been feeling discomfort on his side. Two days later, his lower back started hurting as well, especially when he moves. You noticed that he has not been eating much lately. When you asked him about it, he told you he has been feeling extremely tired and he has no appetite. Today, he called in sick because he has a fever of 38.2°C. As far as you know, he had no accident. Wanting to find out what he could be suffering from, you go to https://www.babylonhealth.com/ask-babylon-chat to check the symptoms he told you about.

Table 30 Description of the situations

Once participants reached the diagnosis page, they were asked to return to the WORD document. They received the instructions:

> In the first part of the evaluation, we would like to know what you instinctively remember from this interaction. Please don't transcribe the answers from the browser: you are not being

graded on your memory; the software is being graded on its rememberability.

Try to use keywords and write what first comes to your mind. This part of the evaluation should not last longer than a couple of minutes. Please don't change your answers after leaving the page.

These are two memory questions (Table 31) they had to answer:

N	Question	Input type
1.	What diagnostics does Babylon give you? Use key words only.	Open-ended
2.	Does Babylon suggest you do something? (such as monitor a symptom)	Open-ended

Table 31 Memory questions after each situation

Afterwards, they are asked to complete the User Experience Questionnaire (UEQ), presented in Figure 13. This is how the task was presented to them:

For the assessment of each situation, please fill out the following questionnaire. The questionnaire consists of pairs of contrasting attributes that may apply to the product. The circles between the attributes represent gradations between the opposites. You can express your agreement with the attributes by clicking on the circle that most closely reflects your impression.

Example:

This response would mean that you rate the application as more annoying than enjoyable.

Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression and don't go back to change your answers.

Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute doesn't apply completely to the particular product. Nevertheless, please click on a circle in every line.

It is your personal opinion that counts. Please remember there is no wrong or right answer!

	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	doesn't meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22
organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

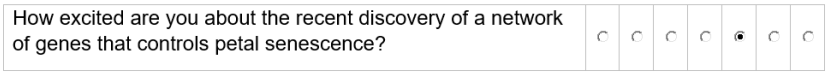
Figure 13 User Experience Questionnaire (UEQ)

In the post-test, administered immediately at the end of the experiment, we ask all participants to evaluate overall the product on a scale of 1 to 7. We then ask them to give us their opinion on current technological developments. They receive these instructions:

Lastly, please complete the following questionnaire.

In this last questionnaire, we ask you to give your opinion about health technologies currently in development. You are asked three close-ended questions per technology. The circles after each question represent gradations between two opposites: “not at all” on the right side and “very much” on the left side. You can express your opinion by clicking on the circle that most closely reflects your opinion.

Example:



This response would mean that you are more excited than not about this discovery.

Please decide spontaneously. Don’t think too long about your decision to make sure that you convey your original impression and don’t go back to change your answers.

Table 32 shows the questions asked during the post-test.

How excited are you about teleHealth (i.e., digitally led health services)?	Scale 1-7	1
How much do you want teleHealth to play a role in the future of medicine?		2
Do you believe teleHealth is an improvement?		3
How excited are you about personalized medicine (i.e., medicine designed for you)?		4
How much do you want personalized medicine to play a role in the future?		5
Do you believe personalized medicine is an improvement?		6
How excited are you about robotic surgery?		7
How much do you want robotic surgery to play a role in the future of medicine?		8
Do you believe robotic surgery is an improvement?		9
How excited are you about artificial intelligence and machine learning in medicine?		10
How much do you want artificial intelligence and machine learning to play a role in the future of medicine?		11
Do you believe artificial intelligence and machine learning in medicine is an improvement?		12
How excited are you about health wearables?		13

How much do you want health wearables to play a role in the future?		14
Do you believe health wearables are an improvement?		15

Table 32 Post-test questionnaire

9.3.2. Corpus data

We conducted the pilot experience on four acquaintances: two males and two females. The females are both from Switzerland. They are 25 and 24 years old. We call them participant A and B respectively. The males are aged 26 and 27. One of them is from Switzerland, the other comes from Belgium. They are participant C and D.

9.3.3. Compatibility issues

Out of four, only one participant successfully completed the WORD document as it was intended. This was due to two main compatibility issues:

1. **Different software:** one of the male participants used LibreOffice. When he reached the end of the experiment, he saved the document and sent it to me. When I opened it, it was completely blank. It did not register properly on his computer either. All his data was lost – I only have my notes left.
2. **Different WORD version:** the two females used WORD on their iOS-based devices. They could not fill out the questionnaire by clicking on the circles.

Beside the compatibility issues, we noticed the document was very heavy, slow and buggy. When exported to PDF, some forms do not work as intended.

Improvements

We will no longer use a WORD document to conduct our experiment. Instead, we will use an online questionnaire maker compatible with our questionnaire, such as Qualtrics or Limesurvey. The software must protect our data, be free and easy to use for our participants.

9.3.4. Formatting issues

Participants were perplexed or made mistakes due to the format of the WORD document. For example, at one point, the document did not specify that the questions they needed to answer were on the following page, so every participant started answering on their current page. Obviously, this was not intended.

Another problem was the placement of vital information. Participant B started registering on the website because she did not notice the log-in information given.

Improvement

We will add more precise instructions regarding where participants are asked to rate Babylon and give their opinion. We will also place the link at the end of the instructions, so participants are not tempted to leave before receiving all the information they need.

9.3.5. Wording issues

During task n°2, participant C did not understand he was meant to use the feature "COVID-19 check" on the website. Instead, he started using the feature tested in situation n°1. Participant B was also unsure about what she should do in that same situation.

Apart from these issues, there were also a couple of mistakes that need to be rectified. There were also a few situations that needed clarifying because crucial information was missing.

Improvement

We clarify situation n°2 as follows:

N	Description of the situation ***clarified*** and rectified
1.	It is high season for your seasonal allergies. Your nose has been ***blocked*** for the past week, so you have been breathing mostly through your mouth. Your jaw hurts a little and you have been feeling stab-like pain in your right ear ***that gradually started*** yesterday. Today, you noticed that you were hearing less from that ear. Your ear feels moist, but nothing has come out. When you look at it, everything seems normal. You have no other symptoms. Your ear bothers you a lot, but you don't want to immediately call the doctor. With only this information in mind, you check your symptoms on the website https://www.babylonhealth.com/ask-babylon-chat
2.	You have been at home for the past three days. Today, from the

	moment you woke up, you have been feeling quite tired, even though you went to bed and woke up at the same hour you usually do. As the day goes by, your body increasingly aches all over. As a precaution, you take your temperature every 2 hours. Last time you checked, you had 37.5°C. After taking your contraceptive pill, you notice that your throat starts to feel funny, like there is something stuck that you cannot quite swallow. It aches slightly. You suspect you got the flu because your windows don't fully close, and cold wind passes through. Outside, the temperature has dropped dramatically since yesterday. At the same time, you work a job that requires frequent contact with other people. No one else in your house has similar symptoms. Even though you are not particularly vulnerable to COVID-19, because you have no underlying conditions, you are a bit worried and you want to check your symptoms. You ***decide to do a COVID-19 check using the*** website https://www.babylonhealth.com/ask-babylon-chat.
3.	Five days ago, your 25-year-old male friend, Jake, told you he has been feeling discomfort on his side. Two days later, his lower back started hurting as well, especially when he moves. You noticed that he has not been eating much lately. When you asked him about it, he told you he has been feeling extremely tired and he has no appetite. Today, he called in sick because he feels very unwell. You learn he has a fever of 38.2°C. As far as you know, he had no accident. Wanting to find out what he could be suffering from, you go to https://www.babylonhealth.com/ask-babylon-chat to check the symptoms he told you about.

Table 33 Modifications to the description of each situation

9.3.6. Timing

Ideally, this experiment lasts around 30 minutes and maximum 45 minutes. In Figure 14, we present the results as follows: on time, overtime, undertime.

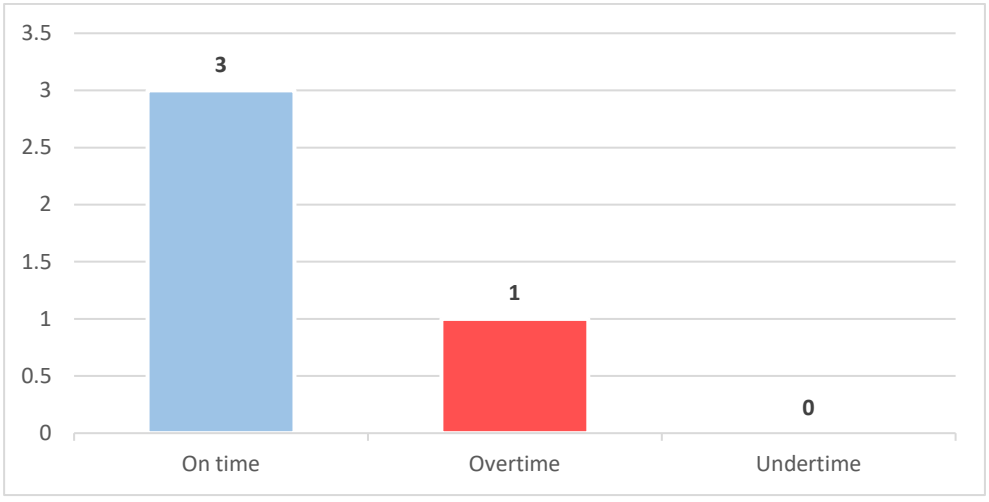


Figure 14 Respect of time restriction

The experiment with participant B lasted 1 hour. She admitted feeling tired at the end of the experiment. Therefore, 45 minutes must not be exceeded. Participant B confessed the results from task 3 might be invalid, because she was no longer paying attention.

Improvement

To complete the three situations, participants require a solid B2, close to C1, English level. Participant B does not have a B2 level in this field of study (medicine) and struggled to understand several items. Therefore, her interaction with Babylon Health took her considerably more time.

To remedy this problem, we decided to reduce the experiment to a single situation and evaluation. In our decision process, we considered the fact that other participants were a bit weary at the end of the experiment because it was too repetitive. Plus, completing three situations causes familiarity with the product: they get quite used to it and lose their excitement. Data from the last questionnaire is thus corrupted.

9.3.7. Results

Content analysis

Having two questions related to memory adds no value to this experiment. During the pilot, we noticed first-hand that even though participants are told they are not being graded on their memory, after they encounter memory-related questions, they concentrate much more on the second task and third task. At the end, they actively try to remember the diagnosis. However, inciting such behaviour is unnecessary and unproductive. Therefore, we will no administer memory question, but instead merely ask participants to pay attention to the results and hint they might be questioned on it, even though they will not be. This should enhance levels of concentration enough.

We also noticed that trust is a crucial concept that is currently understudied in our experiment. Recent studies underline trust as an essential concept when measuring experience of AI-related software. We find the UEQ lacking in that department. Only item 17 ‘secure – not secure’ measures a construct somewhat related to trust. This aspect being insufficiently developed in our experiment, and we do not achieve our goals. Therefore,

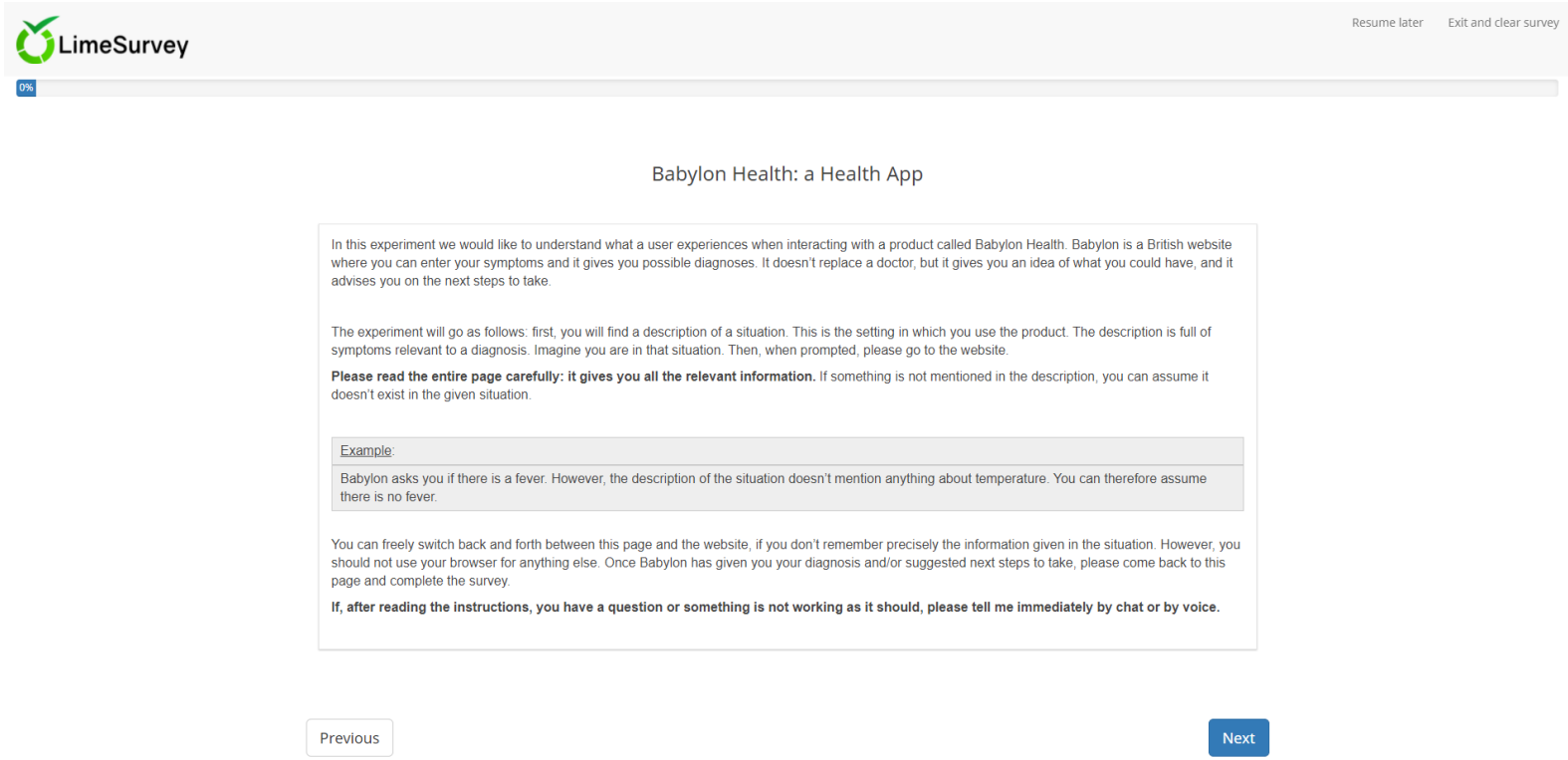
instead of administering the UEQ, we switch to its modular version: the UEQ+ and we add a module concentrating specifically on trust.

Statistical analysis

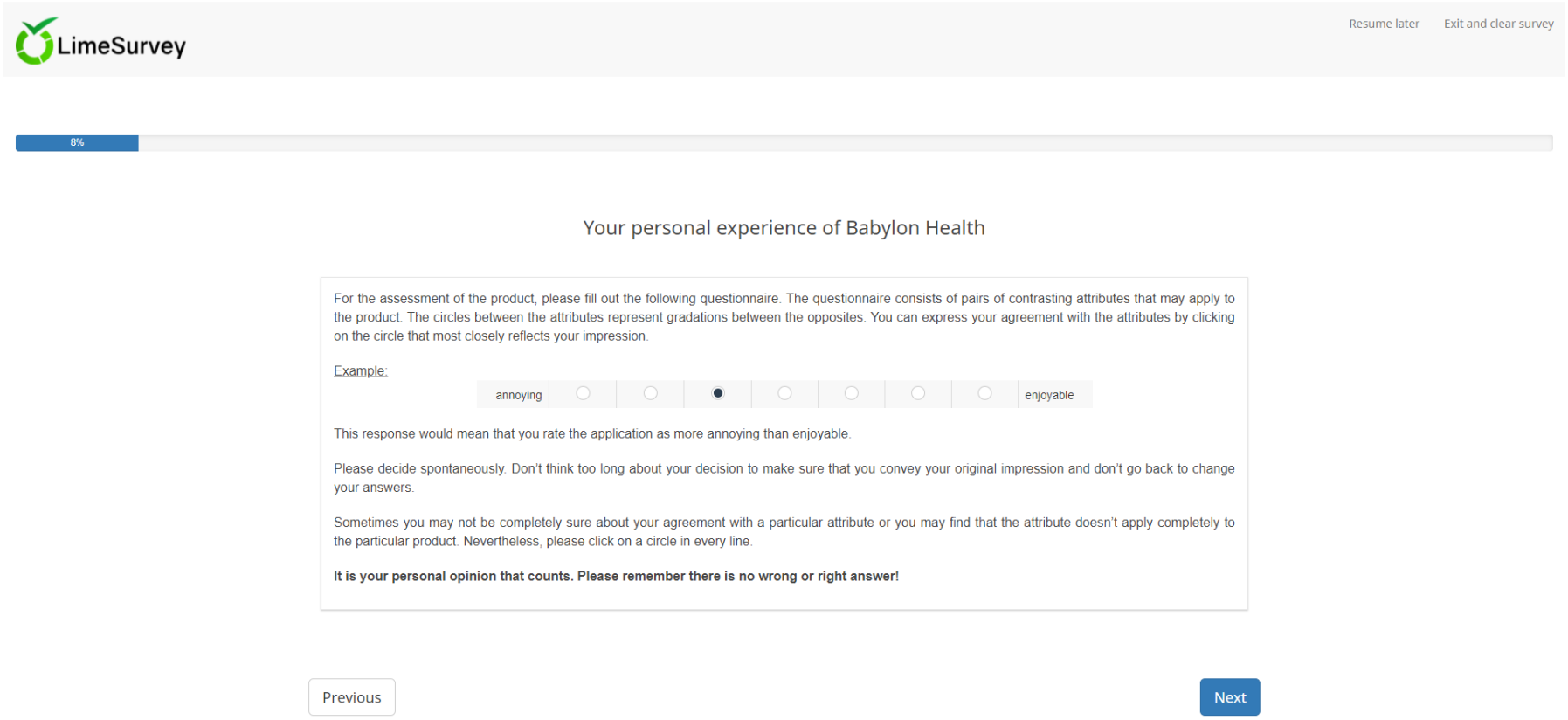
Since data was stock for only 3 of the 4 participants, it is impossible to properly compare the two groups. Plus, the pool being so modest, they do not follow a normal distribution. Finally, this experiment being completely rehandled due to software incompatibility issues and method deficiency, we decided not to further lengthen this paper with fruitless analysis.

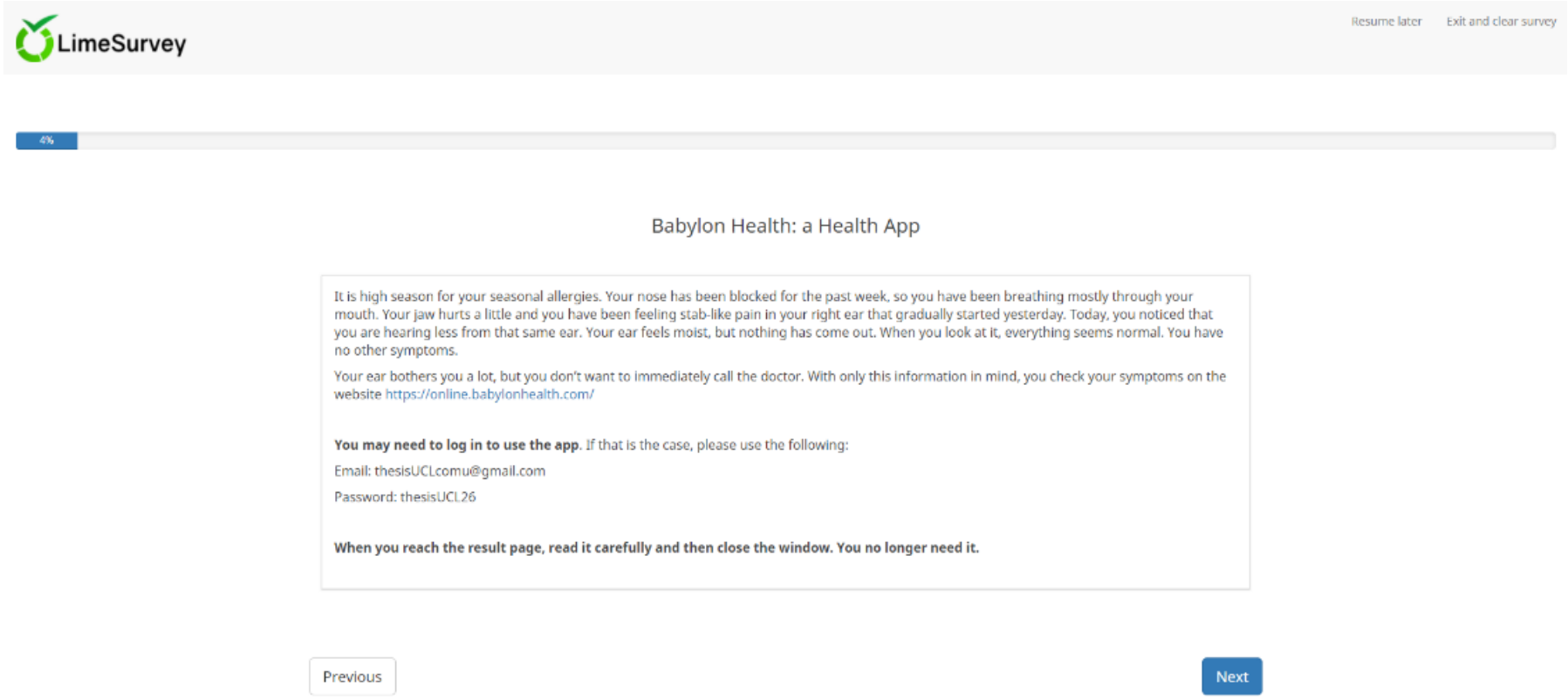
9.4. Survey format in Limesurvey

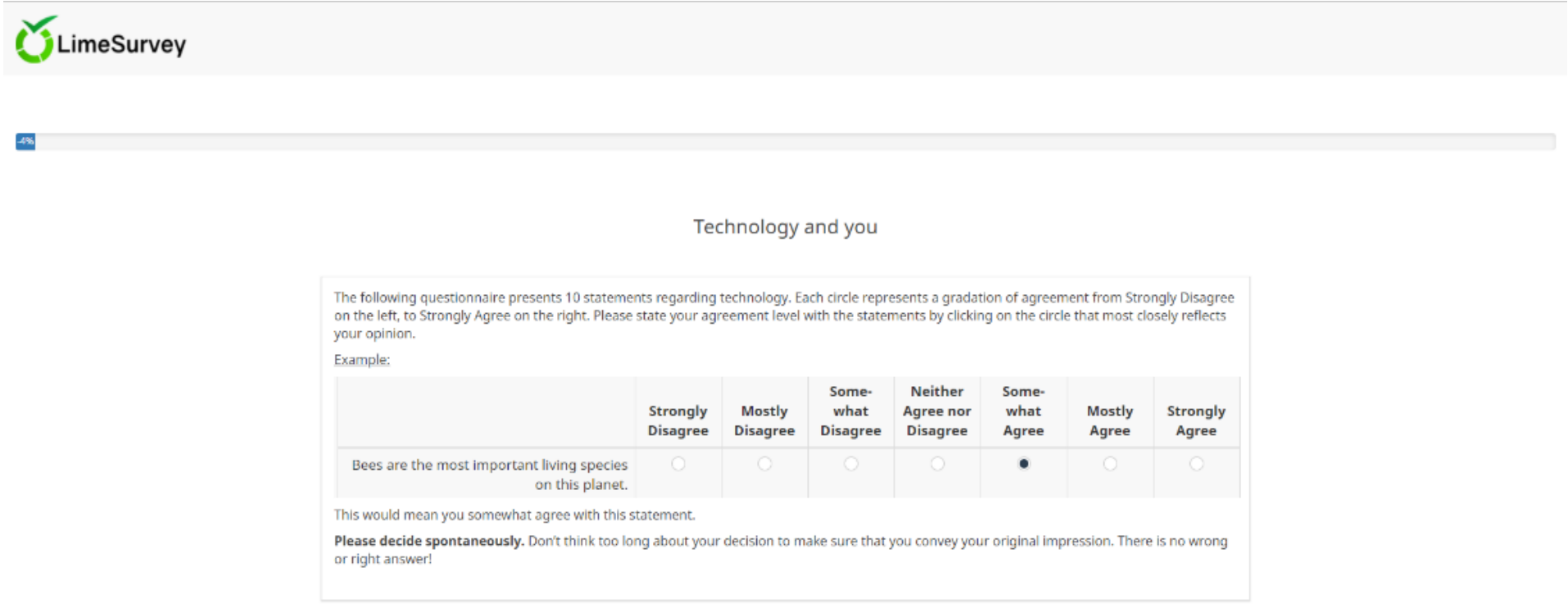
9.4.1. Presentation



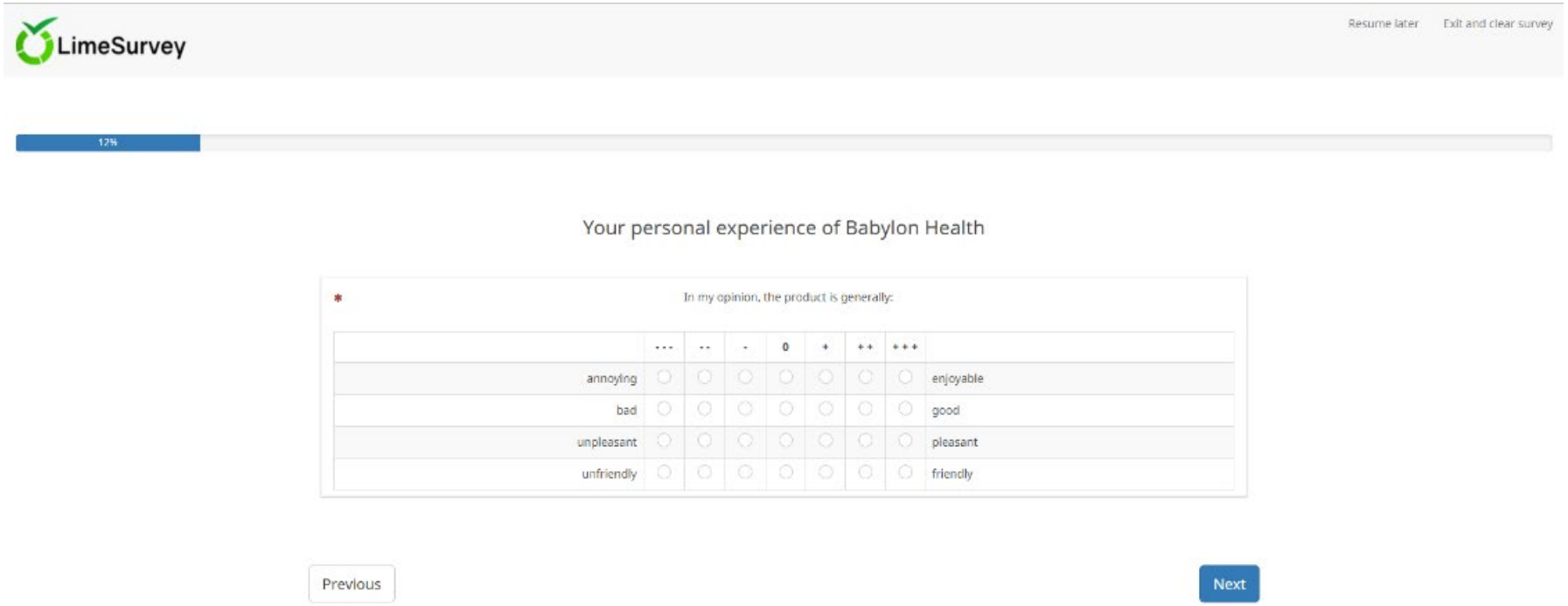
9.4.2. Instructions



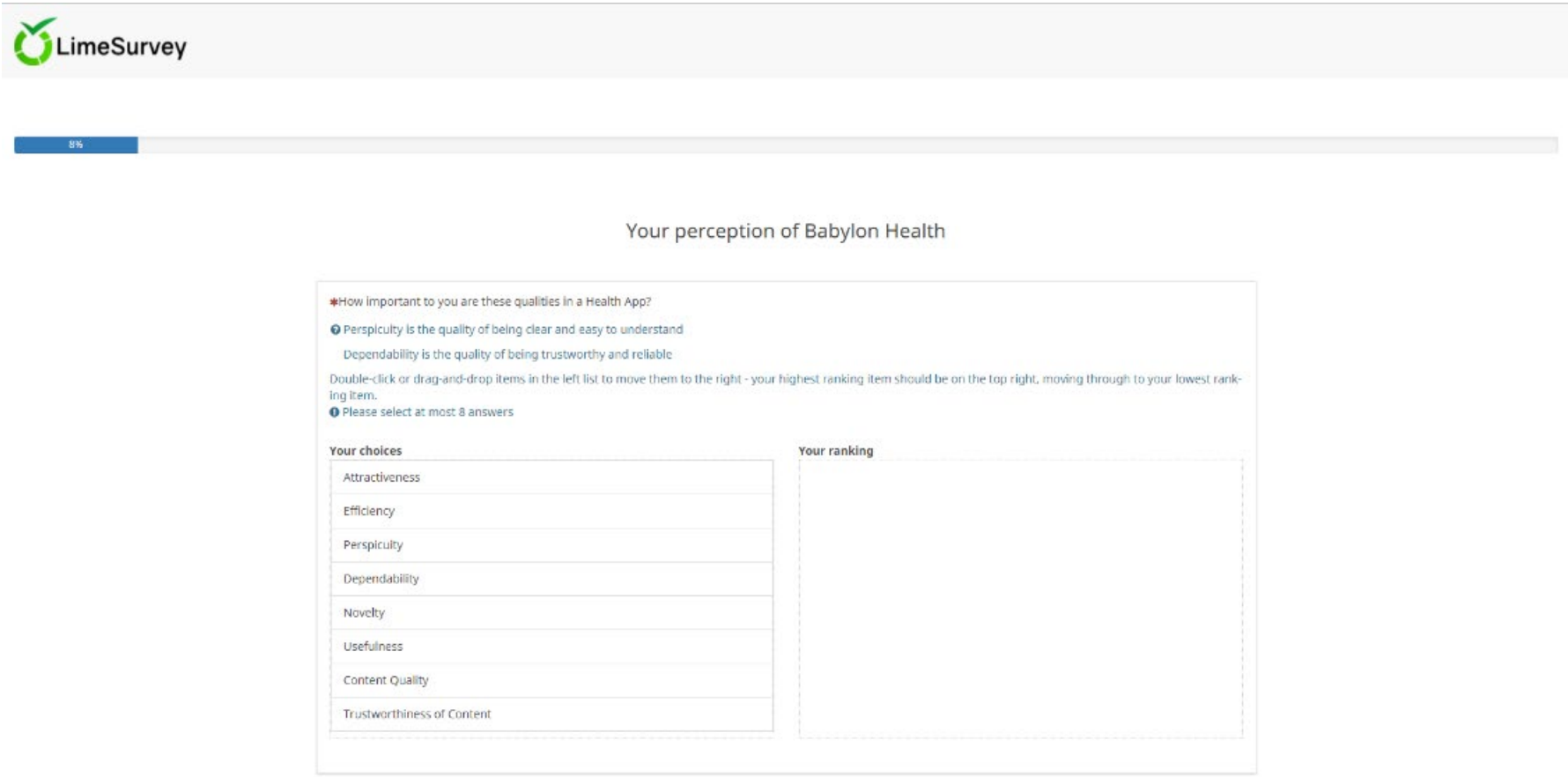

LimeSurvey
Resume later
Exit and clear survey
4%
Babylon Health: a Health App
It is high season for your seasonal allergies. Your nose has been blocked for the past week, so you have been breathing mostly through your mouth. Your jaw hurts a little and you have been feeling stab-like pain in your right ear that gradually started yesterday. Today, you noticed that you are hearing less from that same ear. Your ear feels moist, but nothing has come out. When you look at it, everything seems normal. You have no other symptoms.
Your ear bothers you a lot, but you don't want to immediately call the doctor. With only this information in mind, you check your symptoms on the website https://online.babylonhealth.com/
You may need to log in to use the app. If that is the case, please use the following:
Email: thesisUCLcomu@gmail.com
Password: thesisUCL26
When you reach the result page, read it carefully and then close the window. You no longer need it.
Previous
Next


LimeSurvey
4%
Technology and you
The following questionnaire presents 10 statements regarding technology. Each circle represents a gradation of agreement from Strongly Disagree on the left, to Strongly Agree on the right. Please state your agreement level with the statements by clicking on the circle that most closely reflects your opinion.
Example:
Strongly Disagree
Mostly Disagree
Some-what Disagree
Neither Agree nor Disagree
Some-what Agree
Mostly Agree
Strongly Agree
Bees are the most important living species on this planet.
This would mean you somewhat agree with this statement.
Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression. There is no wrong or right answer!

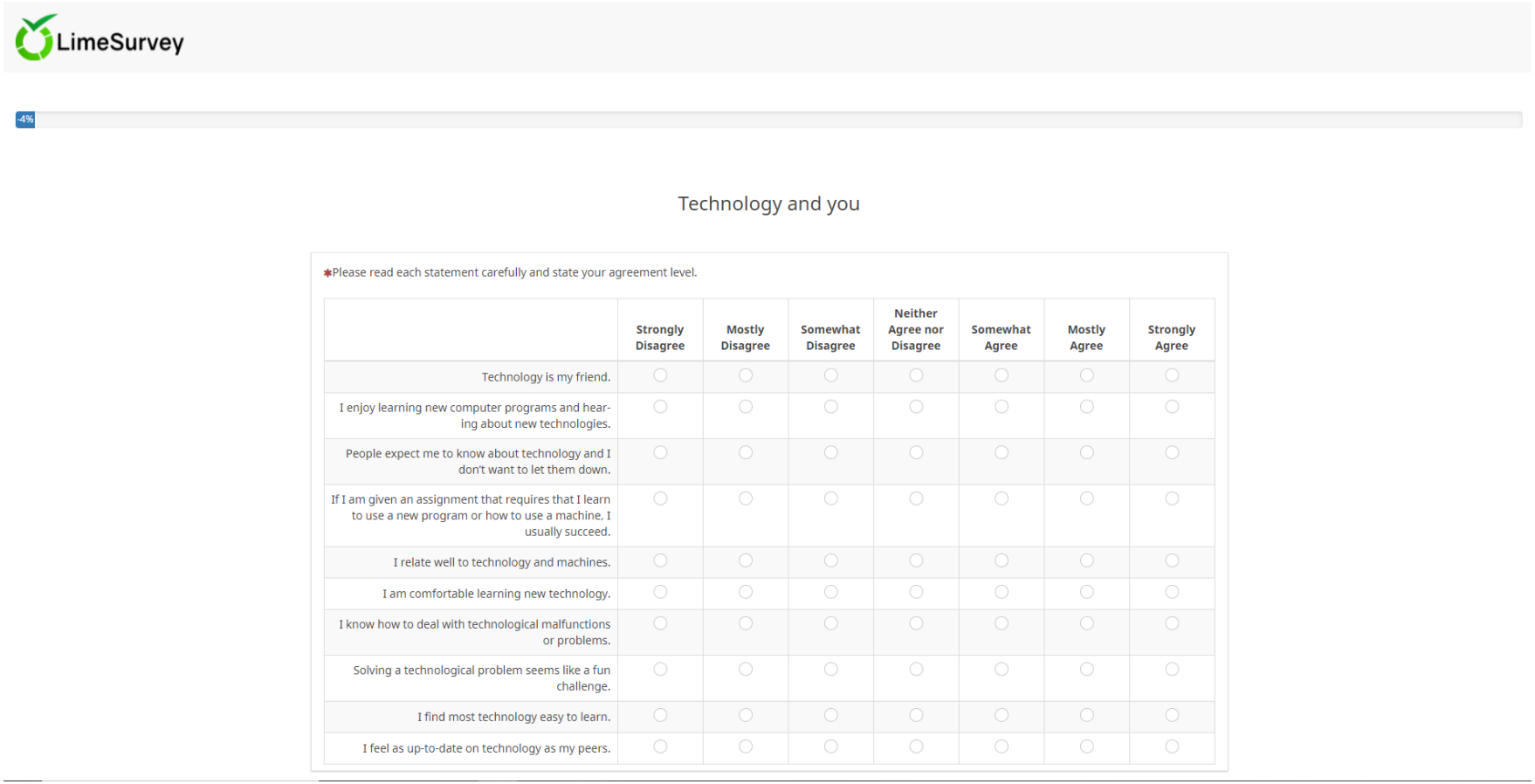
UEQ+ questions



9.4.3. Ranking question



9.4.4. Attitude towards technology questions



LimeSurvey

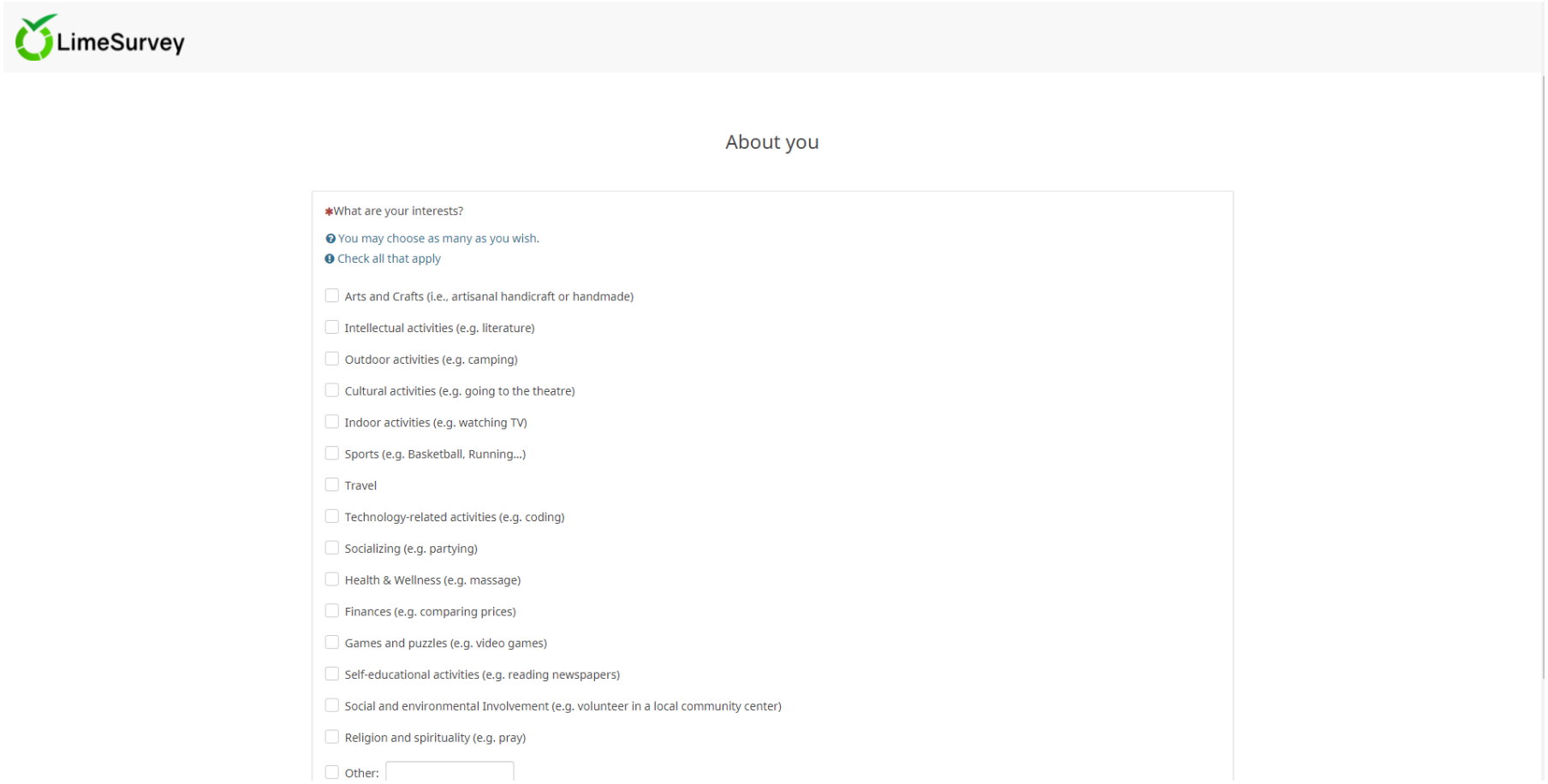
4%

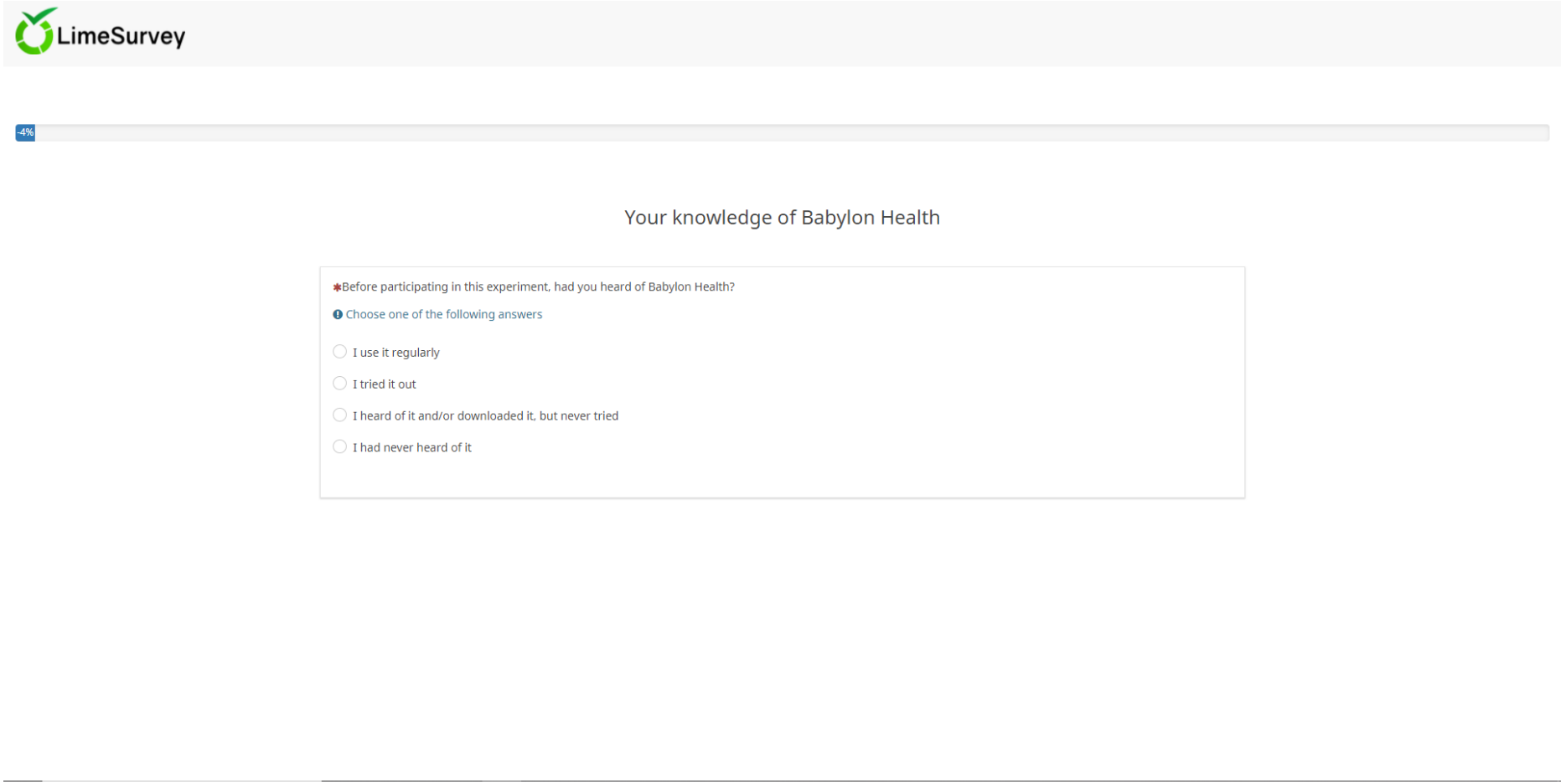
Technology and you

*Please read each statement carefully and state your agreement level.

	Strongly Disagree	Mostly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Mostly Agree	Strongly Agree
Technology is my friend.	○	○	○	○	○	○	○
I enjoy learning new computer programs and hearing about new technologies.	○	○	○	○	○	○	○
People expect me to know about technology and I don't want to let them down.	○	○	○	○	○	○	○
If I am given an assignment that requires that I learn to use a new program or how to use a machine, I usually succeed.	○	○	○	○	○	○	○
I relate well to technology and machines.	○	○	○	○	○	○	○
I am comfortable learning new technology.	○	○	○	○	○	○	○
I know how to deal with technological malfunctions or problems.	○	○	○	○	○	○	○
Solving a technological problem seems like a fun challenge.	○	○	○	○	○	○	○
I find most technology easy to learn.	○	○	○	○	○	○	○
I feel as up-to-date on technology as my peers.	○	○	○	○	○	○	○

9.4.5. Demographics format




LimeSurvey
-4%
Your knowledge of Babylon Health
*Before participating in this experiment, had you heard of Babylon Health?
Choose one of the following answers
I use it regularly
I tried it out
I heard of it and/or downloaded it, but never tried
I had never heard of it

9.5. Results

9.5.1. Mean, Variance and Standard Deviation

		No Treatment			Treatment		
	Scale	Mean	Var.	Std. dev.	Mean	Var.	Std. dev.
Mix population – LCOMU2812	Attractiveness	1.36	1.17	1.07	1.35	1.00	0.99
	Efficiency	1.61	1.82	1.34	1.08	1.92	1.37
	Perspicuity	2.16	0.83	0.90	1.93	1.81	1.33
	Dependability	1.64	1.17	1.07	1.08	2.07	1.42
	Usefulness	1.45	1.37	1.16	1.30	1.65	1.27
	Novelty	0.98	1.79	1.32	0.75	3.17	1.76
	Trustworthiness of Content	1.66	0.70	0.82	1.50	1.18	1.07
	Content Quality	2.20	0.68	0.81	1.90	0.76	0.86
Female population	Attractiveness	1.53	0.92	0.95	1.47	1.06	1.01
	Efficiency	1.70	1.65	1.27	1.64	1.27	1.11
	Perspicuity	2.25	0.71	0.83	2.11	1.64	1.26
	Dependability	1.73	1.13	1.05	1.44	1.68	1.28
	Usefulness	1.18	2.15	1.45	1.25	2.25	1.48
	Novelty	0.65	1.87	1.35	1.14	2.41	1.53
	Trustworthiness of Content	1.45	0.82	0.89	1.67	1.43	1.18
	Content Quality	2.18	0.66	0.80	1.94	0.91	0.94
Mix population – all participants	Attractiveness	5.47	0.69	0.83	5.37	0.35	0.59
	Efficiency	5.55	1.23	1.11	5.27	0.72	0.85
	Perspicuity	6.22	0.41	0.64	6.14	1.06	1.03
	Dependability	5.62	0.90	0.95	5.18	0.75	0.87
	Usefulness	5.32	1.21	1.10	5.34	0.94	0.97
	Novelty	4.85	1.32	1.15	4.86	2.35	1.53
	Trustworthiness of Content	5.57	0.42	0.64	5.70	0.75	0.87
	Content Quality	6.20	0.67	0.61	5.63	0.59	0.77

9.5.2. Descriptive Statistics by treatment, all participants included

	Treatment			Statistic	Std. Error
Attractiveness	No Treatment	Mean		5.4667	.21390
		95% Confidence Interval for Mean	Lower Bound	5.0079	
			Upper Bound	5.9254	
		5% Trimmed Mean		5.4491	
		Median		5.7500	
		Variance		.686	
		Std. Deviation		.82844	
		Minimum		4.25	
		Maximum		7.00	
		Range		2.75	
		Interquartile Range		1.25	
		Skewness		.108	.580
		Kurtosis		-.865	1.121
	Treatment	Mean		5.3750	.15887
		95% Confidence Interval for Mean	Lower Bound	5.0318	
			Upper Bound	5.7182	
		5% Trimmed Mean		5.3611	
		Median		5.3750	
		Variance		.353	
		Std. Deviation		.59445	
		Minimum		4.50	
		Maximum		6.50	
		Range		2.00	
		Interquartile Range		.88	
		Skewness		.461	.597
		Kurtosis		-.595	1.154
Efficiency	No Treatment	Mean		5.5500	.28681
		95% Confidence Interval for Mean	Lower Bound	4.9348	
			Upper Bound	6.1652	
		5% Trimmed Mean		5.6111	

		Median		5.5000	
		Variance		1.234	
		Std. Deviation		1.11082	
		Minimum		3.00	
		Maximum		7.00	
		Range		4.00	
		Interquartile Range		1.25	
		Skewness		-.649	.580
		Kurtosis		.749	1.121
	Treatment	Mean		5.2679	.22615
		95% Confidence Interval for Mean	Lower Bound	4.7793	
			Upper Bound	5.7564	
		5% Trimmed Mean		5.2837	
		Median		5.1250	
		Variance		.716	
		Std. Deviation		.84617	
		Minimum		3.50	
		Maximum		6.75	
		Range		3.25	
		Interquartile Range		1.25	
		Skewness		-.058	.597
		Kurtosis		.452	1.154
Perspicuity	No Treatment	Mean		6.2167	.16523
		95% Confidence Interval for Mean	Lower Bound	5.8623	
			Upper Bound	6.5711	
		5% Trimmed Mean		6.2407	
		Median		6.2500	
		Variance		.410	
		Std. Deviation		.63994	
		Minimum		5.00	
		Maximum		7.00	
		Range		2.00	
		Interquartile Range		1.00	
		Skewness		-.390	.580
		Kurtosis		-.811	1.121
	Treatment	Mean		6.1429	.27451
		95% Confidence Interval for	Lower Bound	5.5498	

		Mean	Upper Bound	6.7359	
		5% Trimmed Mean		6.2143	
		Median		6.3750	
		Variance		1.055	
		Std. Deviation		1.02711	
		Minimum		4.00	
		Maximum		7.00	
		Range		3.00	
		Interquartile Range		1.31	
		Skewness		-1.078	.597
		Kurtosis		.284	1.154
Dependability	No Treatment	Mean		5.6167	.24503
		95% Confidence Interval for Mean	Lower Bound	5.0911	
			Upper Bound	6.1422	
		5% Trimmed Mean		5.6296	
		Median		5.7500	
		Variance		.901	
		Std. Deviation		.94900	
		Minimum		4.00	
		Maximum		7.00	
		Range		3.00	
		Interquartile Range		1.50	
		Skewness		-.038	.580
		Kurtosis		-.715	1.121
	Treatment	Mean		5.1786	.23209
		95% Confidence Interval for Mean	Lower Bound	4.6772	
			Upper Bound	5.6800	
		5% Trimmed Mean		5.1429	
		Median		4.7500	
		Variance		.754	
		Std. Deviation		.86840	
		Minimum		4.00	
		Maximum		7.00	
		Range		3.00	
		Interquartile Range		1.13	
		Skewness		.820	.597
		Kurtosis		-.036	1.154

Usefulness	No Treatment	Mean		5.3167	.28501
		95% Confidence Interval for Mean	Lower Bound	4.7054	
			Upper Bound	5.9280	
		5% Trimmed Mean		5.3241	
		Median		5.2500	
		Variance		1.218	
		Std. Deviation		1.10384	
		Minimum		3.50	
		Maximum		7.00	
		Range		3.50	
		Interquartile Range		1.25	
		Skewness		.036	.580
		Kurtosis		-.744	1.121
	Treatment	Mean		5.3393	.25892
		95% Confidence Interval for Mean	Lower Bound	4.7799	
			Upper Bound	5.8986	
		5% Trimmed Mean		5.3353	
		Median		5.2500	
		Variance		.939	
		Std. Deviation		.96878	
		Minimum		3.75	
		Maximum		7.00	
		Range		3.25	
		Interquartile Range		1.44	
		Skewness		.054	.597
		Kurtosis		-.560	1.154
Novelty	No Treatment	Mean		4.8500	.29661
		95% Confidence Interval for Mean	Lower Bound	4.2138	
			Upper Bound	5.4862	
		5% Trimmed Mean		4.8750	
		Median		4.5000	
		Variance		1.320	
		Std. Deviation		1.14876	
		Minimum		3.00	
		Maximum		6.25	
		Range		3.25	

		Interquartile Range		2.25	
		Skewness		-.057	.580
		Kurtosis		-1.654	1.121
	Treatment	Mean		4.8571	.40997
		95% Confidence Interval for Mean	Lower Bound	3.9715	
			Upper Bound	5.7428	
		5% Trimmed Mean		4.9107	
		Median		5.1250	
		Variance		2.353	
		Std. Deviation		1.53396	
		Minimum		2.00	
		Maximum		6.75	
		Range		4.75	
		Interquartile Range		2.56	
		Skewness		-.599	.597
		Kurtosis		-.615	1.154
Trustworthiness of Content	No Treatment	Mean		5.5667	.16631
		95% Confidence Interval for Mean	Lower Bound	5.2100	
			Upper Bound	5.9234	
		5% Trimmed Mean		5.5463	
		Median		5.5000	
		Variance		.415	
		Std. Deviation		.64411	
		Minimum		4.50	
		Maximum		7.00	
		Range		2.50	
		Interquartile Range		.75	
		Skewness		.238	.580
		Kurtosis		.843	1.121
	Treatment	Mean		5.4107	.22938
		95% Confidence Interval for Mean	Lower Bound	4.9152	
			Upper Bound	5.9063	
		5% Trimmed Mean		5.4147	
		Median		5.2500	
		Variance		.737	
		Std. Deviation		.85826	

		Minimum		4.00	
		Maximum		6.75	
		Range		2.75	
		Interquartile Range		1.44	
		Skewness		.315	.597
		Kurtosis		-.743	1.154
Content Quality	No Treatment	Mean		6.2000	.15660
		95% Confidence Interval for Mean	Lower Bound	5.8641	
			Upper Bound	6.5359	
		5% Trimmed Mean		6.2222	
		Median		6.2500	
		Variance		.368	
		Std. Deviation		.60651	
		Minimum		5.00	
		Maximum		7.00	
		Range		2.00	
		Interquartile Range		1.00	
		Skewness		-.251	.580
		Kurtosis		-.672	1.121
	Treatment	Mean		5.9107	.18458
		95% Confidence Interval for Mean	Lower Bound	5.5119	
			Upper Bound	6.3095	
		5% Trimmed Mean		5.9147	
		Median		5.8750	
		Variance		.477	
		Std. Deviation		.69065	
		Minimum		4.75	
		Maximum		7.00	
		Range		2.25	
		Interquartile Range		1.06	
		Skewness		.076	.597
		Kurtosis		-.709	1.154

9.5.3. Tests of Normality

Mix population – LCOMU2812 by treatment group

	Group	Statistic	df	Sig.
Attractiveness	No Treatment	.934	11	.451
	Treatment	.929	10	.442
Efficiency	No Treatment	.899	11	.178
	Treatment	.945	10	.608
Perspicuity	No Treatment	.944	11	.565
	Treatment	.850	10	**.**059
Dependability	No Treatment	.895	11	.159
	Treatment	.842	10	**.046**
Usefulness	No Treatment	.955	11	.714
	Treatment	.989	10	.995
Novelty	No Treatment	.841	11	**.032**
	Treatment	.904	10	.244
Content Quality	No Treatment	.884	11	.119
	Treatment	.895	10	.195
Trustworthiness of Content	No Treatment	.969	11	.874
	Treatment	.974	10	.925

Table 34 Test of Normality by treatment – LCOMU2812 population

Mix population– LCOMU2812 by treatment and by gender

	Gender	No Treatment			Treatment		
		Statistic	df	Sig.	Statistic	df	Sig.
Attractiveness	Male	.894	5	.375	.914	5	.490
	Female	.832	6	.111	.925	5	.564
Efficiency	Male	.974	5	.898	.710	5	**.012**
	Female	.874	6	.242	.961	5	.814
Perspicuity	Male	.931	5	.603	.775	5	**.050**
	Female	.979	6	.945	.836	5	.154
Dependability	Male	.945	5	.703	.552	5	**.000**
	Female	.900	6	.376	.913	5	.485
Usefulness	Male	.907	5	.451	.925	5	.563
	Female	.994	6	.996	.958	5	.795
Novelty	Male	.836	5	.154	.931	5	.603
	Female	.868	6	.218	.857	5	.218
Content Quality	Male	.944	5	.692	.684	5	**.006**
	Female	.840	6	.131	.915	5	.501
Trustworthiness of Content	Male	.888	5	.346	.943	5	.685
	Female	.957	6	.794	.989	5	.976

Table 35 Test of Normality by gender and by treatment – LCOMU2812 population

Female population by treatment group

	Group	Statistic	df	Sig.
Attractiveness	No Treatment	.958	10	.759
	Treatment	.966	9	.860
Efficiency	No Treatment	.907	10	.264
	Treatment	.967	9	.871
Perspicuity	No Treatment	.923	10	.380
	Treatment	.756	9	**.006**
Dependability	No Treatment	.920	10	.355
	Treatment	.977	9	.947
Usefulness	No Treatment	.952	10	.693
	Treatment	.945	9	.634
Novelty	No Treatment	.930	10	.452
	Treatment	.885	9	.179
Content Quality	No Treatment	.917	10	.332
	Treatment	.917	9	.369
Trustworthiness of Content	No Treatment	.936	10	.508
	Treatment	.943	9	.611

Table 36 Shapiro-Wilk test by treatment group

Female population by Moment users knew it was an AI-system

	Moment users knew it was an AI-system	**Shapiro-Wilk**		
		Statistic	df	Sig.
Attractiveness	Knew before the question was asked	.955	12	.708
	Did not know	.984	7	.978
Efficiency	Knew before the question was asked	.945	12	.568
	Did not know	.967	7	.873
Perspicuity	Knew before the question was asked	.816	12	.014
	Did not know	.868	7	.179
Dependability	Knew before the question was asked	.946	12	.576
	Did not know	.940	7	.636
Usefulness	Knew before the question was asked	.910	12	.210
	Did not know	.930	7	.549
Novelty	Knew before the question was asked	.960	12	.780
	Did not know	.981	7	.963
Trustworthiness of Content	Knew before the question was asked	.944	12	.547
	Did not know	.902	7	.342
Content Quality	Knew before the question was asked	.927	12	.353
	Did not know	.945	7	.686

Mix population– all participants by group

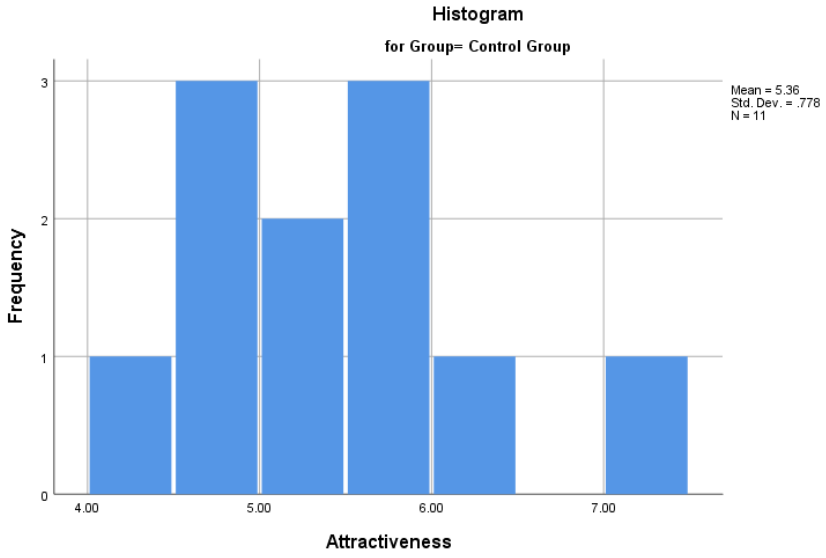
	Treatment	Shapiro-Wilk		
		Statistic	df	Sig.
Attractiveness	No Treatment	.953	15	.576
	Treatment	.955	14	.644
Efficiency	No Treatment	.923	15	.215
	Treatment	.960	14	.720
Perspicuity	No Treatment	.930	15	.275
	Treatment	.808	14	**.006**
Dependability	No Treatment	.917	15	.173
	Treatment	.918	14	.207
Usefulness	No Treatment	.953	15	.578
	Treatment	.974	14	.925
Novelty	No Treatment	.889	15	.065
	Treatment	.927	14	.273
Trustworthiness of Content	No Treatment	.957	15	.632
	Treatment	.921	14	.226
Content Quality	No Treatment	.934	15	.312
	Treatment	.968	14	.848

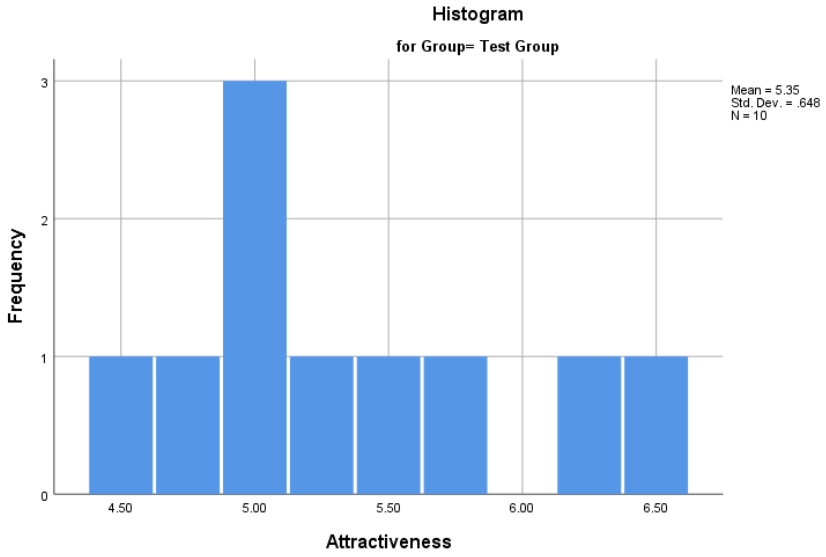
Mix population– all participants by gender

	Gender	Shapiro-Wilk		
		Statistic	df	Sig.
Attractiveness	Male	.907	10	.264
	Female	.963	19	.632
Efficiency	Male	.947	10	.633
	Female	.948	19	.372
Perspicuity	Male	.891	10	.173
	Female	.828	19	**.003**
Dependability	Male	.841	10	**.045**
	Female	.949	19	.378
Usefulness	Male	.901	10	.226
	Female	.934	19	.209
Novelty	Male	.948	10	.646
	Female	.955	19	.474
Trustworthiness of Content	Male	.891	10	.174
	Female	.969	19	.757
Content Quality	Male	.968	10	.872
	Female	.946	19	.331

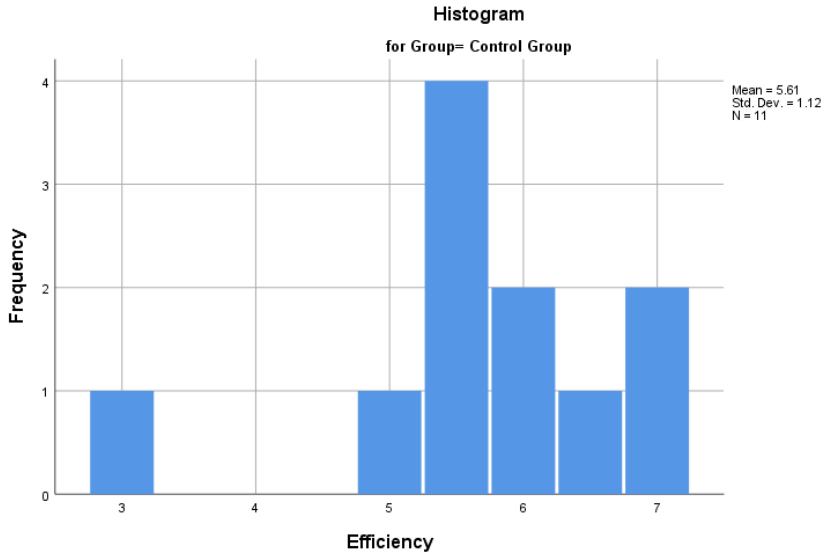
Histograms Mix population by group – LCOMU2812

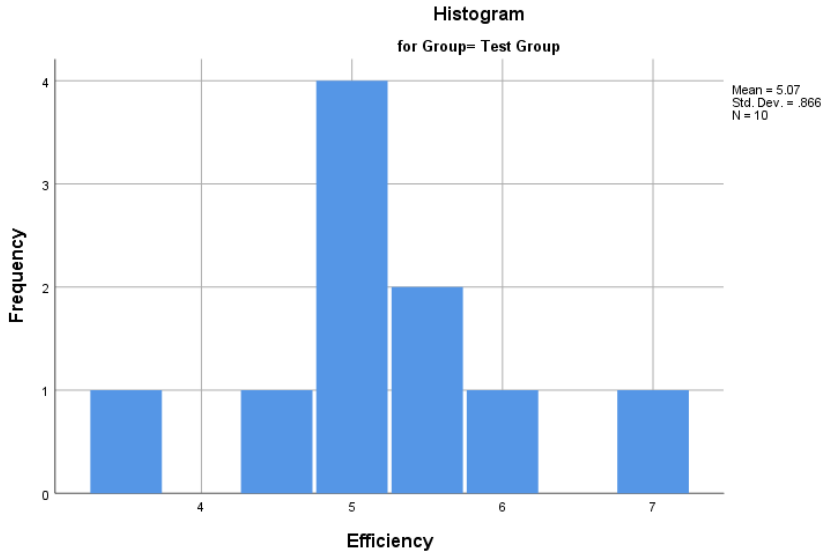
A. Attractiveness



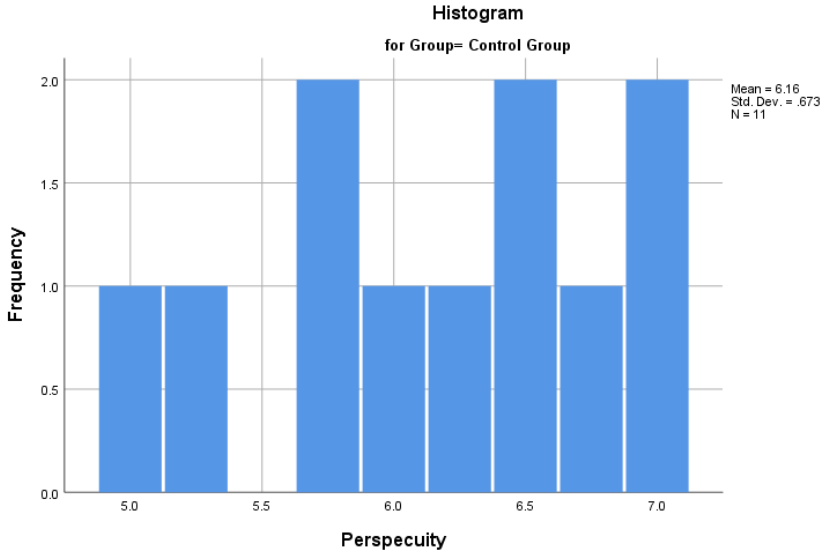


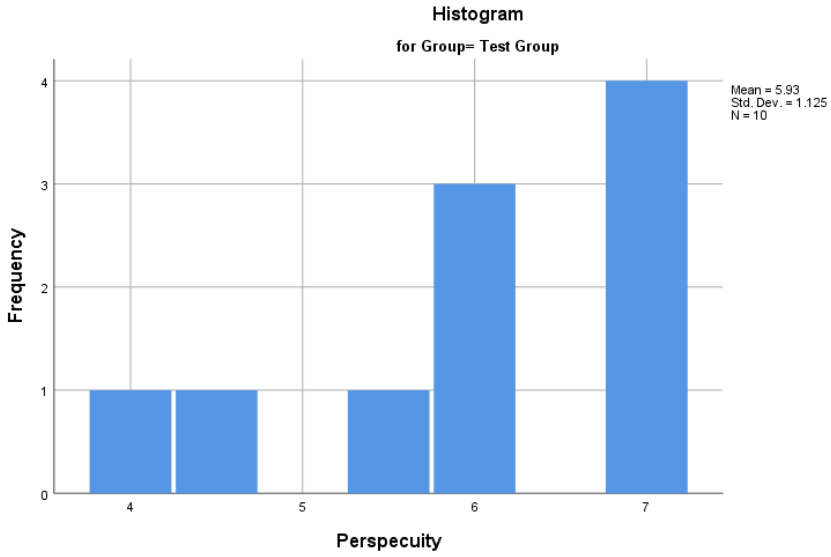
B. Efficiency



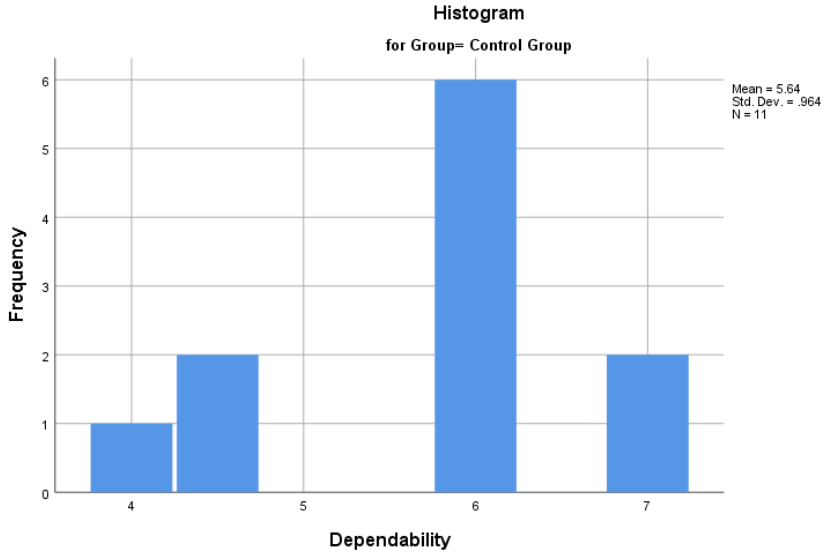


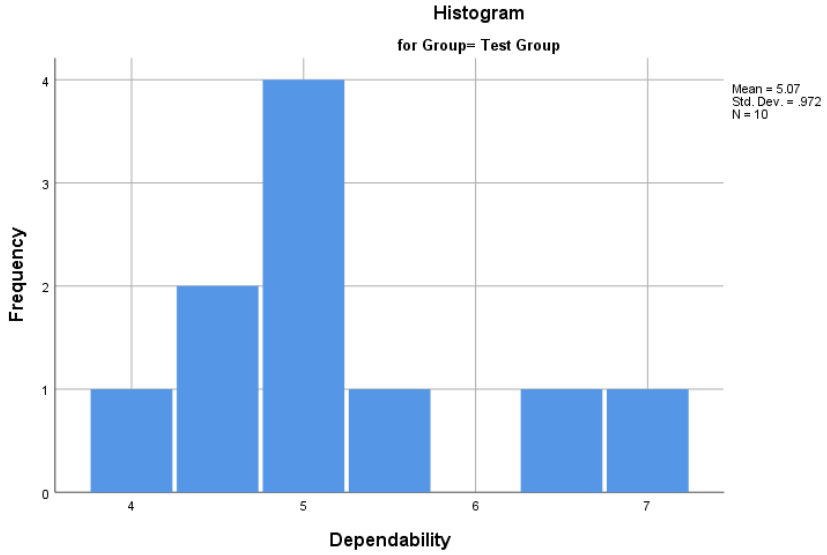
C. Perspicuity



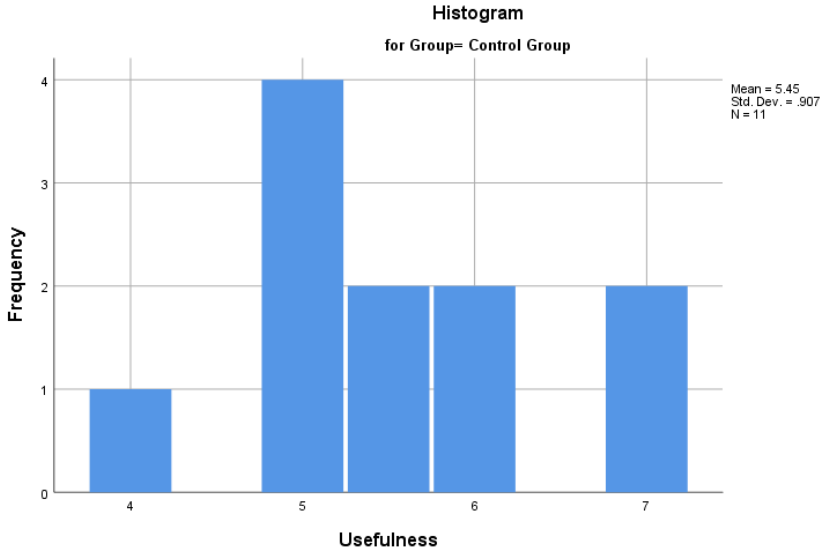


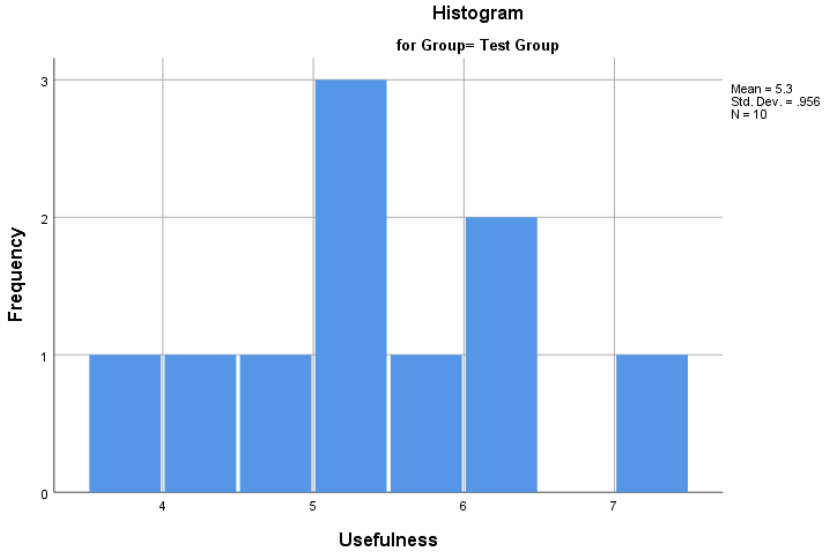
D. Dependability



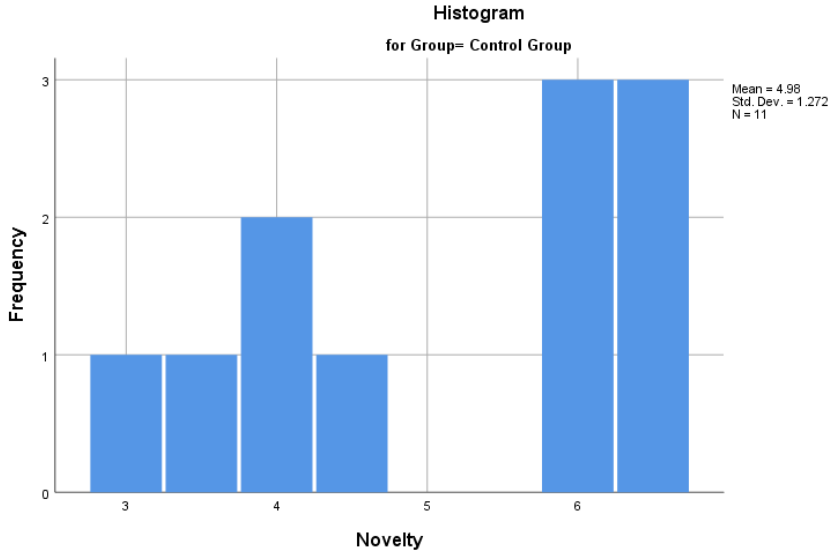


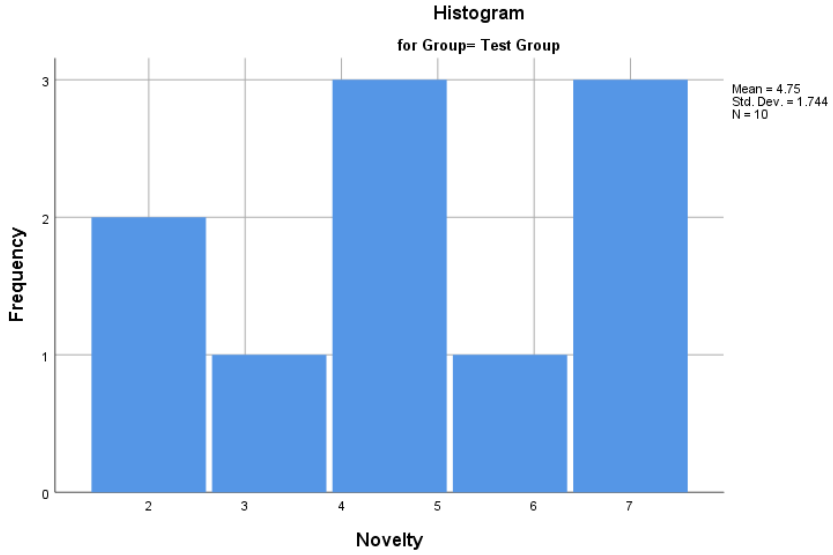
E. Usefulness



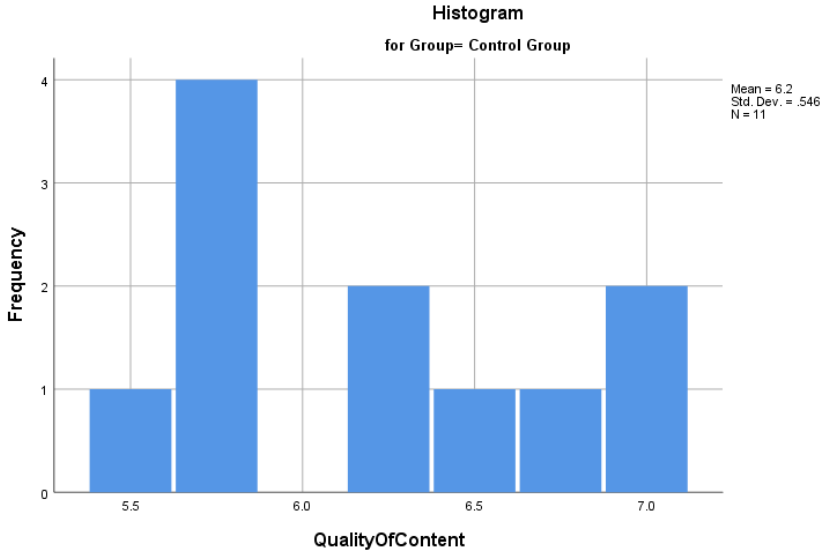


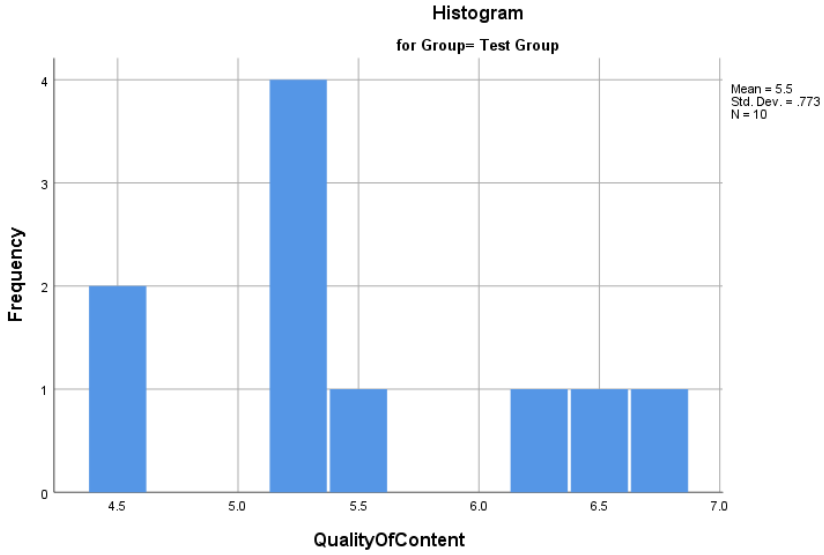
F. Novelty


Histogram
for Group= Control Group
Mean = 4.98
Std. Dev. = 1.272
N = 11
Frequency
Novelty
0
1
2
3
3
4
5
6

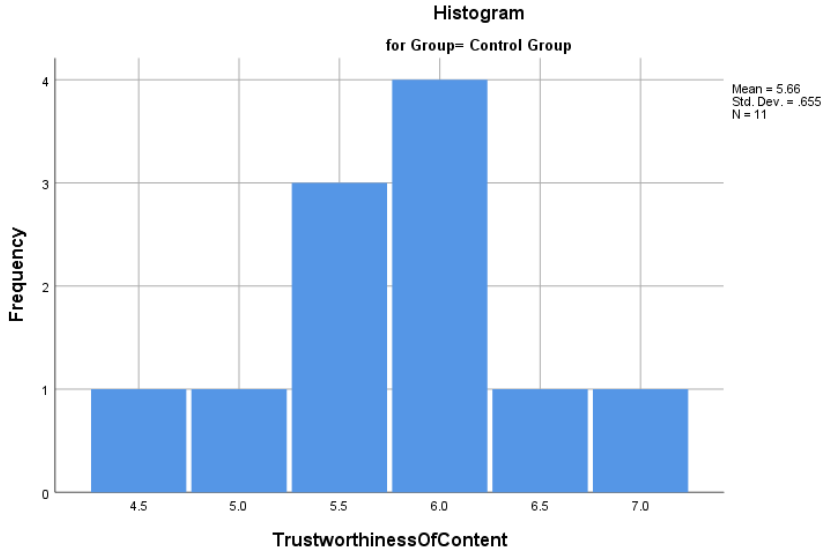

Histogram
for Group= Test Group
Mean = 4.75
Std. Dev. = 1.744
N = 10
Frequency
Novelty
0
1
2
3
2
3
4
5
6
7

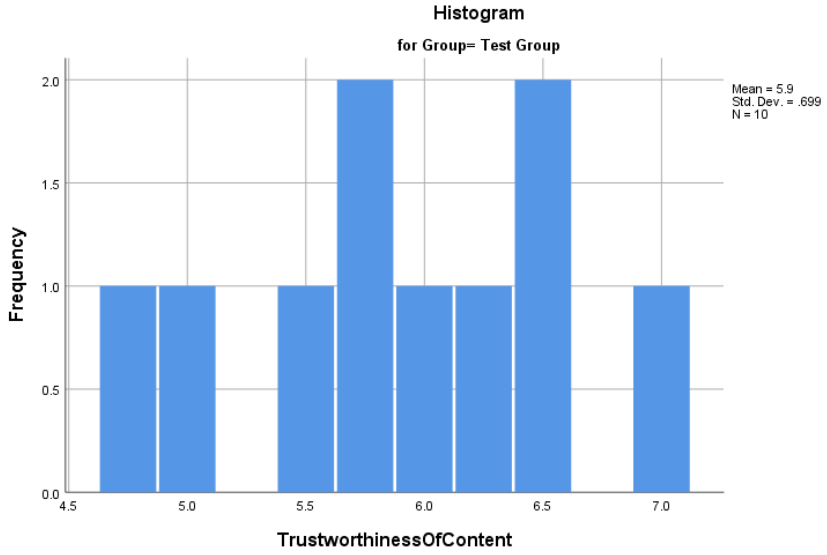
G. Content Quality





H. Trustworthiness of Content





Female population: Histograms by variable and group

A. Attractiveness

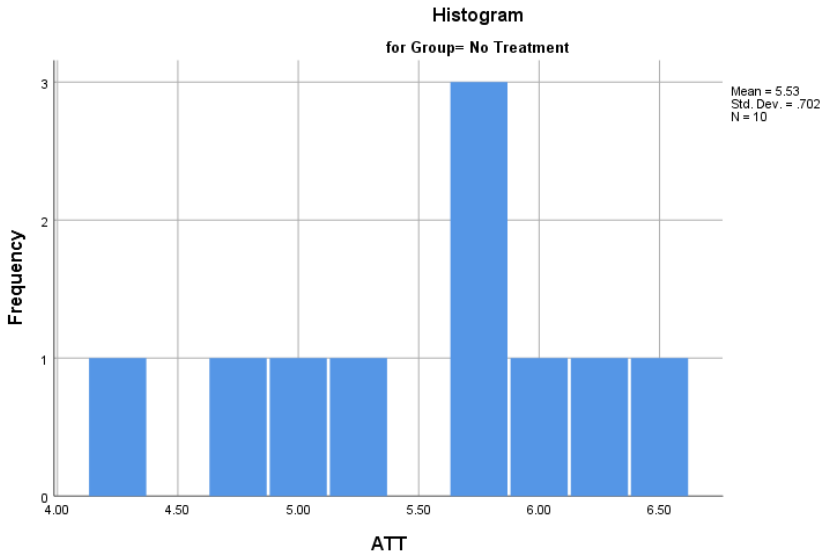


Table 37 Attractiveness histogram for Female population, no treatment group

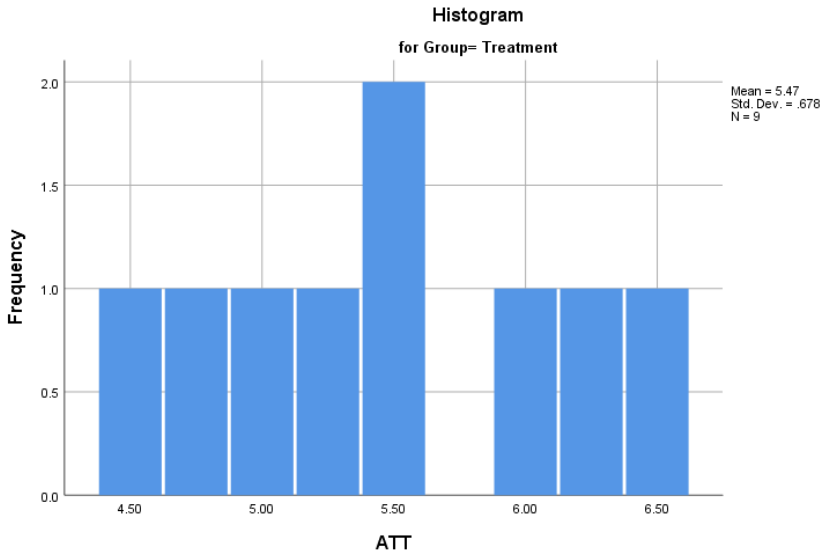


Table 38 Attractiveness histogram for Female population, treatment group

B. Efficiency

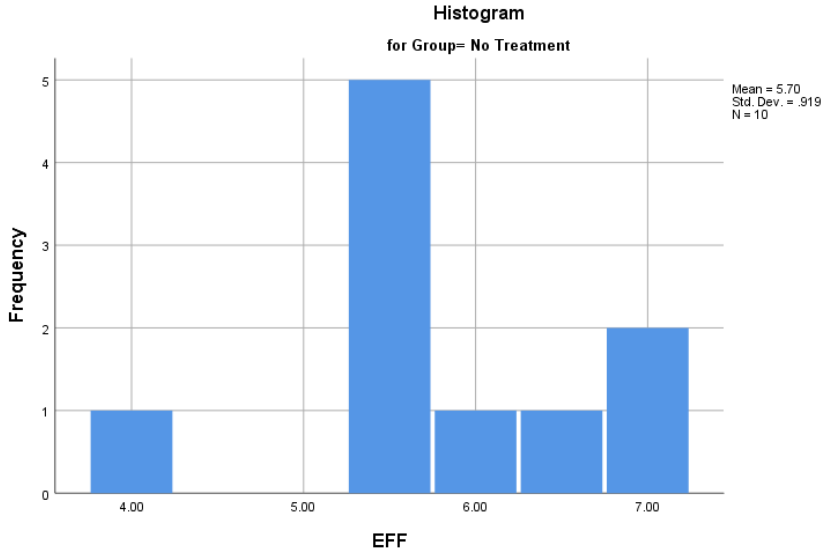


Table 39 Efficiency histogram for Female population, no treatment group

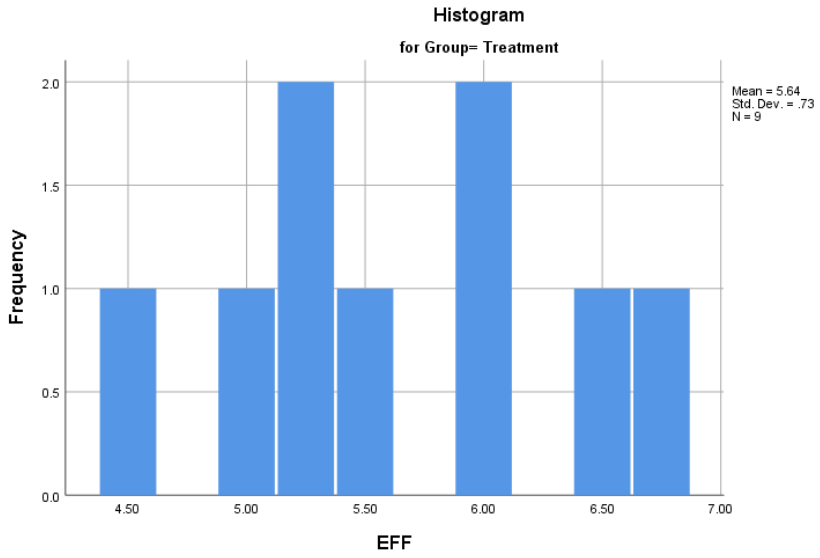


Table 40 Efficiency histogram for Female population, treatment group

C. Perspicuity

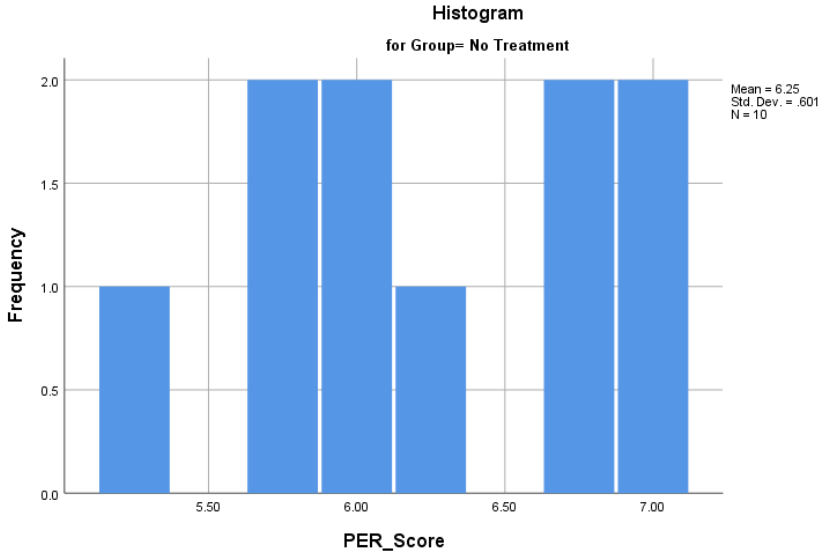


Table 41 Perspicuity histogram for Female population, no treatment group

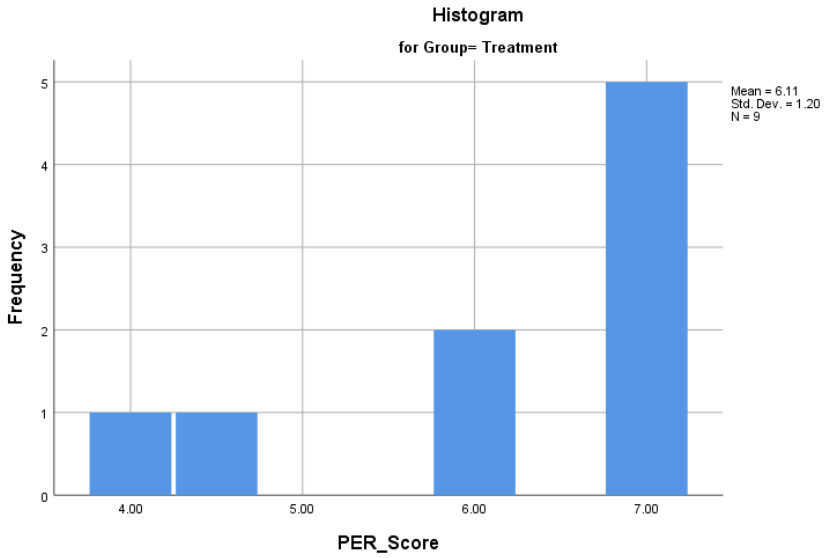


Table 42 Perspicuity histogram for Female population, treatment group

D. Dependability

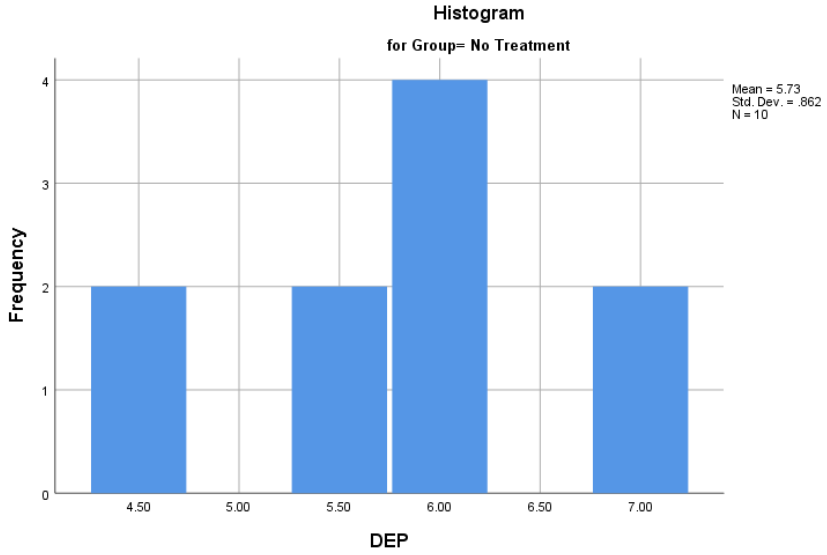


Table 43 Dependability histogram for Female population, no treatment group

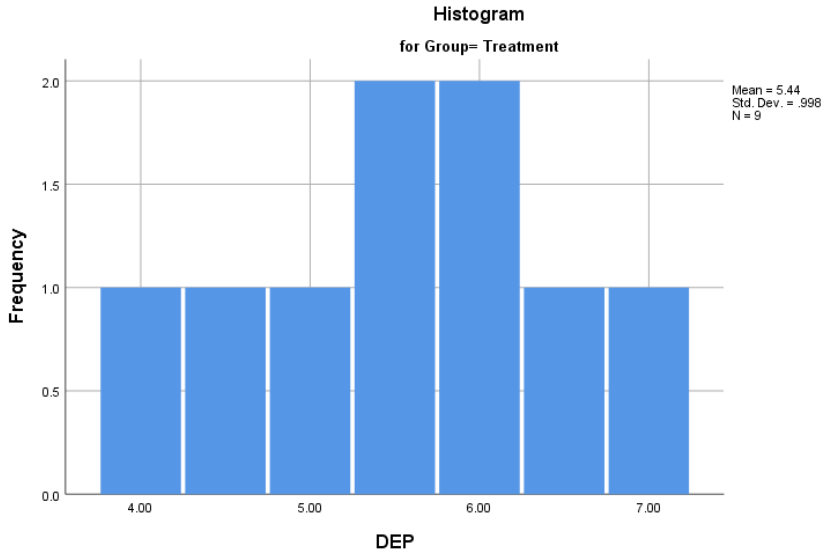


Table 44 Dependability histogram for Female population, treatment group

E. Usefulness

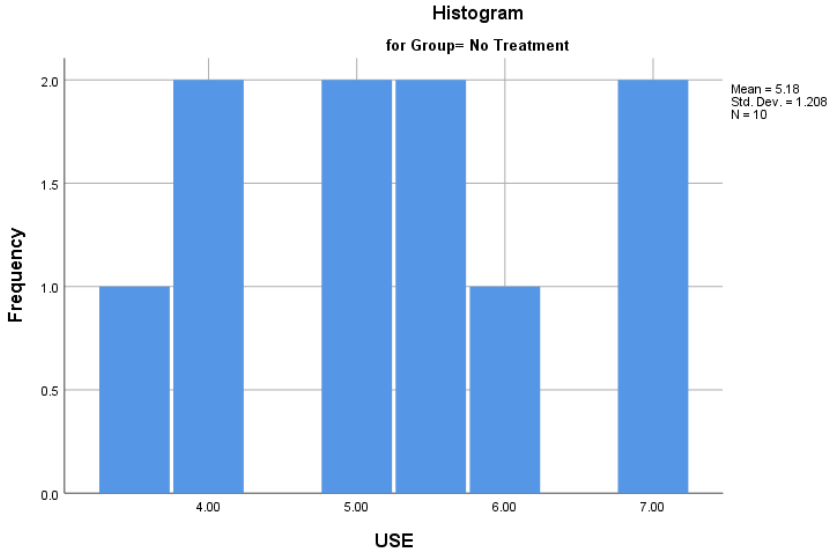


Table 45 Usefulness histogram for Female population, no treatment group

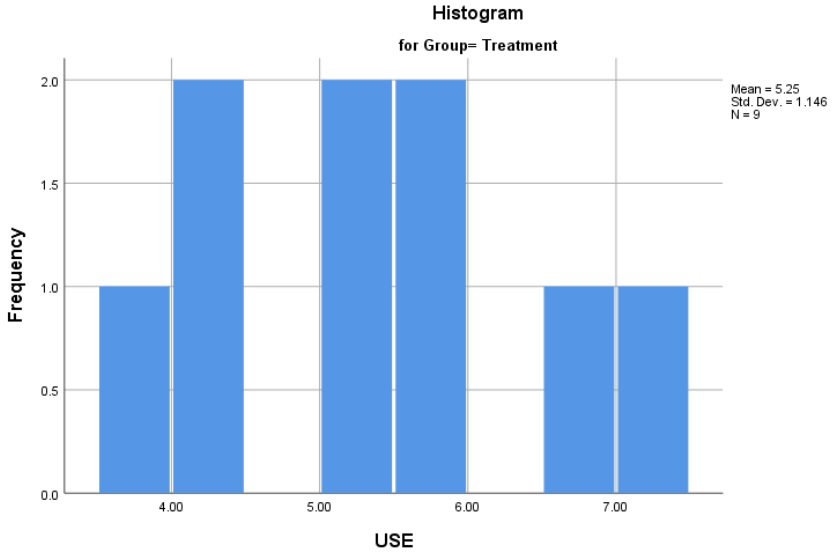


Table 46 Usefulness histogram for Female population, treatment group

F. Novelty

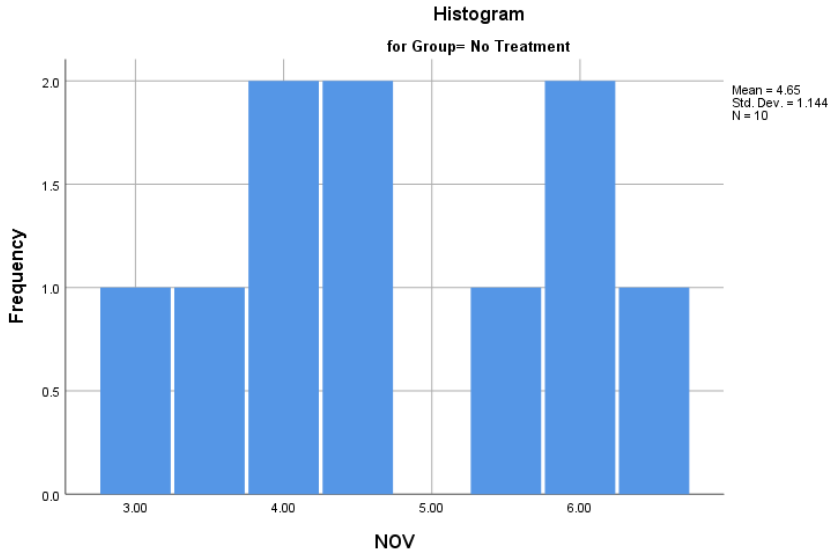


Table 47 Novelty histogram for Female population, no treatment group

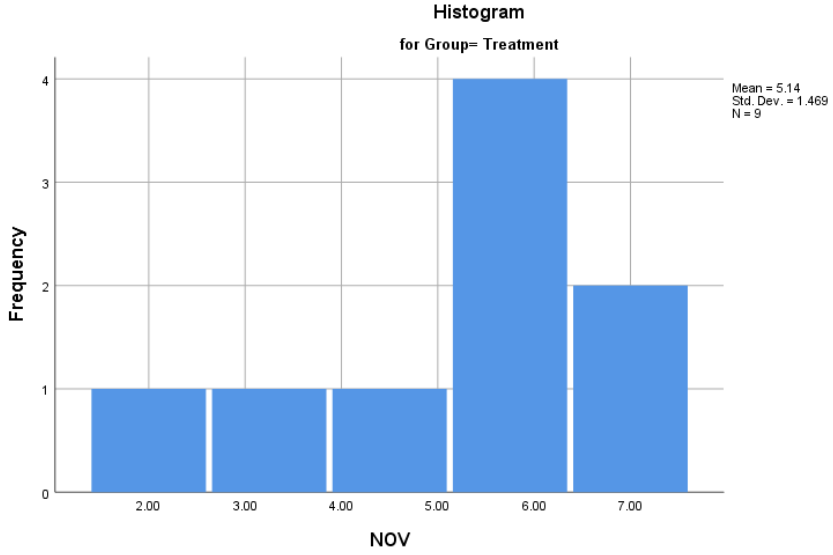


Table 48 Novelty histogram for Female population, treatment group

G. Trustworthiness of Content

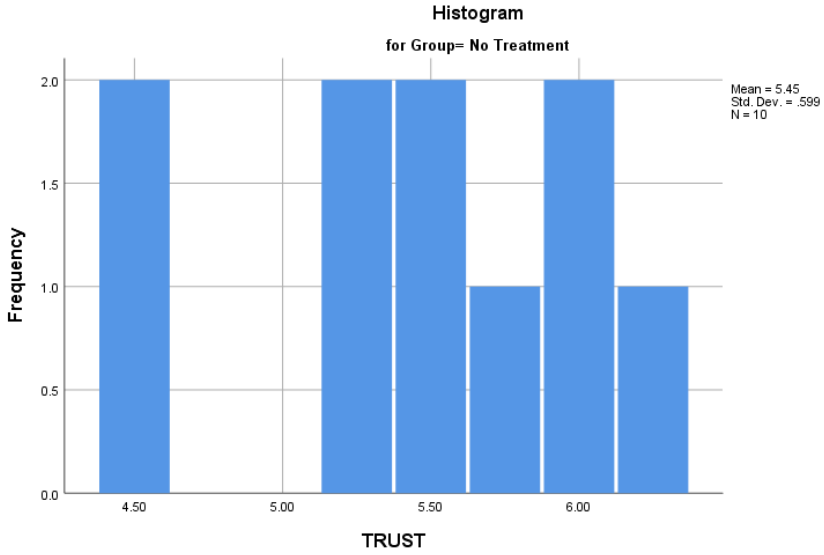


Table 49 Trustworthiness of Content histogram for Female population, no treatment group

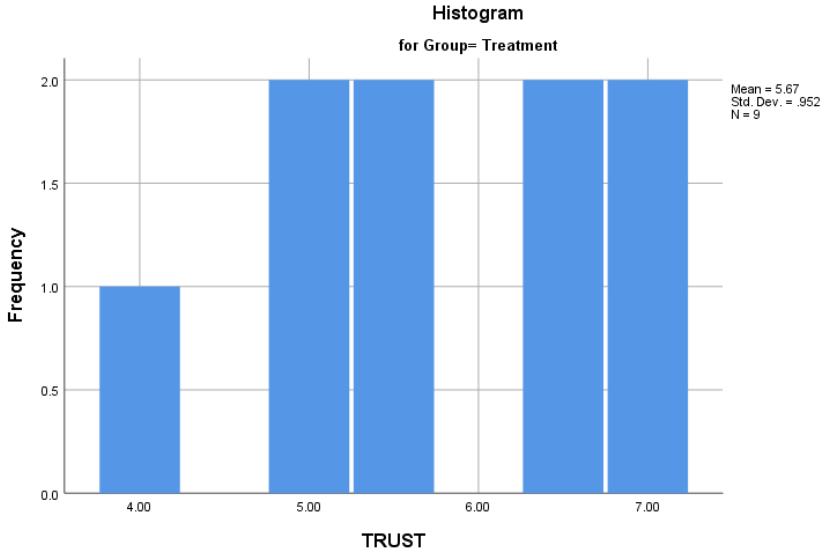


Table 50 Trustworthiness of Content histogram for Female population, treatment group

H. Content Quality

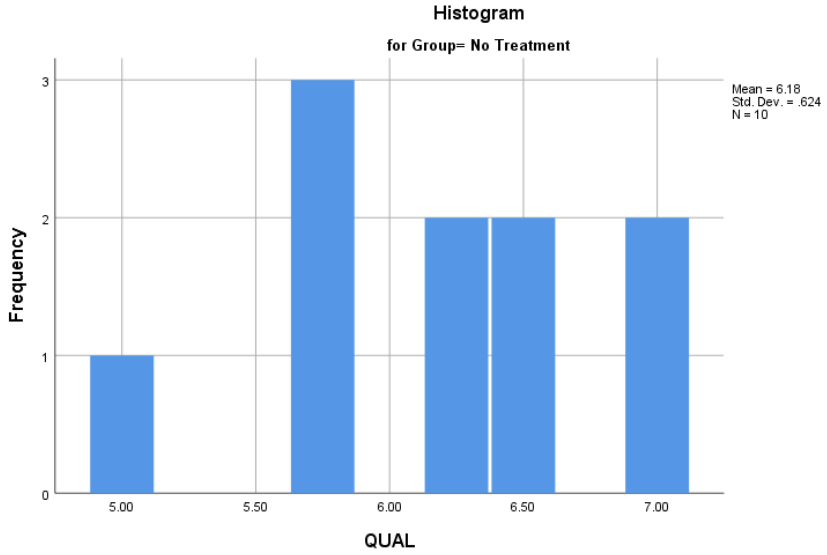


Table 51 Content Quality histogram for Female population, no treatment group

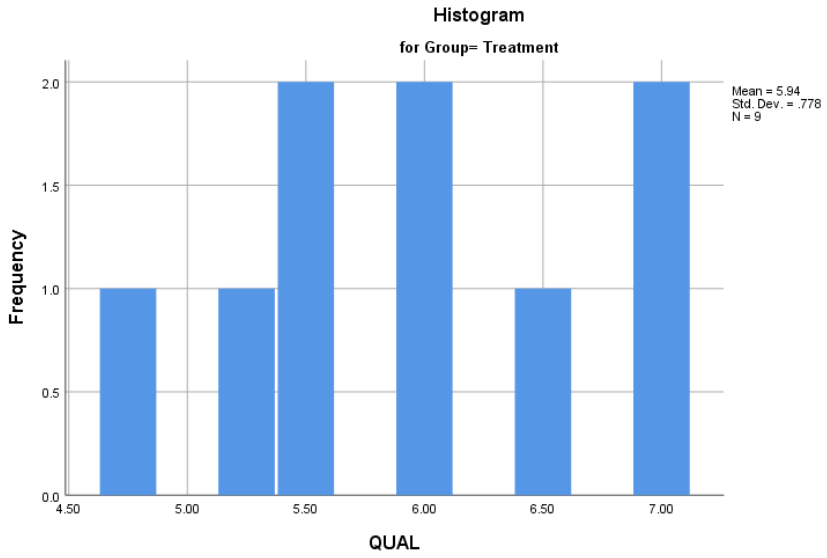


Table 52 Content Quality histogram for Female population, treatment group

9.5.4. Independent Sample t-test/Mann-Whitney U test

No Treatment group - difference between genders

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Attractiveness	Equal variances assumed	7.532	.023	-.050	9	.961	-.02500	.49621	-1.14751	1.09751
	Equal variances not assumed			-.047	5.018	.965	-.02500	.53483	-1.39835	1.34835
Efficiency	Equal variances assumed	2.120	.179	-.981	9	.352	-.667	.680	-2.204	.870
	Equal variances not assumed			-.918	5.374	.398	-.667	.726	-2.496	1.162
Perspecuity	Equal variances assumed	.318	.587	-.039	9	.970	-.017	.430	-.989	.956
	Equal variances not assumed			-.038	7.812	.971	-.017	.438	-1.031	.998

Dependability	Equal variances assumed	1.419	.264	-.724	9	.487	-.433	.598	-1.787	.920
	Equal variances not assumed			-.698	6.851	.508	-.433	.621	-1.908	1.042
Usefulness	Equal variances assumed	.011	.920	.466	9	.652	.267	.572	-1.027	1.561
	Equal variances not assumed			.469	8.825	.650	.267	.569	-1.023	1.557
Novelty	Equal variances assumed	1.364	.273	.629	9	.545	.500	.795	-1.298	2.298
	Equal variances not assumed			.641	8.998	.537	.500	.780	-1.264	2.264
Content Quality	Equal variances assumed	.256	.625	.240	9	.816	.083	.347	-.702	.869
	Equal variances not assumed			.235	7.724	.820	.083	.355	-.739	.906
Trustworthiness of Content	Equal variances assumed	.007	.936	.632	9	.543	.258	.409	-.666	1.183
	Equal variances not assumed			.621	7.912	.552	.258	.416	-.703	1.219

Treatment group – difference between genders

Group Statistics					
	Gender	N	Mean	Std. Deviation	Std. Error Mean
Attractiveness	Male	5	5.3500	1.12639	.50374
	Female	6	5.3750	.44017	.17970
Efficiency	Male	5	5.2500	1.50000	.67082
	Female	6	5.9167	.68313	.27889
Perspicuity	Male	5	6.1500	.78262	.35000
	Female	6	6.1667	.64550	.26352
Dependability	Male	5	5.4000	1.18057	.52797
	Female	6	5.8333	.80104	.32702
Usefulness	Male	5	5.6000	.91173	.40774
	Female	6	5.3333	.97040	.39616
Novelty	Male	5	5.2500	1.17260	.52440
	Female	6	4.7500	1.41421	.57735
Content Quality	Male	5	6.2500	.63738	.28504
	Female	6	6.1667	.51640	.21082
Trustworthiness of Content	Male	5	5.8000	.73739	.32977
	Female	6	5.5417	.62082	.25345

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Attractiveness	Equal variances assumed	4.966	.056	-.712	8	.497	-.30000	.42131	-1.27154	.67154
	Equal variances not assumed			-.712	5.780	.504	-.30000	.42131	-1.34048	.74048
Usefulness	Equal variances assumed	1.506	.255	.639	8	.540	.40000	.62550	-1.04241	1.84241
	Equal variances not assumed			.639	5.796	.547	.40000	.62550	-1.14368	1.94368
Novelty	Equal variances assumed	.004	.954	-.705	8	.501	-.80000	1.13523	-3.41785	1.81785
	Equal variances not assumed			-.705	7.884	.501	-.80000	1.13523	-3.42456	1.82456
Trustworthiness	Equal variances	.876	.377	-	8	.836	-.10000	.46771	-1.17853	.97853

of Content	assumed			.214						
	Equal variances not assumed			-.214	6.924	.837	-.10000	.46771	-1.20843	1.00843

Null Hypothesis	Test	Sig.	Decision
The distribution of Efficiency is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.095	Retain the null hypothesis.
The distribution of Perspicuity is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.841	Retain the null hypothesis.
The distribution of Dependability is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.690	Retain the null hypothesis.
The distribution of Content Quality is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.841	Retain the null hypothesis.
The distribution of Trustworthiness of Content is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.016	Reject the null hypothesis.

Mix population – LCOMU2812 between treatments

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Attractiveness	Equal variances assumed	.380	.545	.043	19	.966	.01364	.31412	-.64383	.67111
	Equal variances not assumed			.044	18.874	.966	.01364	.31129	-.63820	.66547
Efficiency	Equal variances assumed	.305	.587	1.223	19	.236	.53864	.44039	-.38310	1.46038
	Equal variances not assumed			1.239	18.562	.231	.53864	.43488	-.37303	1.45030
Perspicuity	Equal variances assumed	2.102	.163	.585	19	.565	.23409	.39989	-.60290	1.07108
	Equal variances not assumed			.572	14.442	.576	.23409	.40948	-.64165	1.10983
Usefulness	Equal variances assumed	.003	.957	.380	19	.708	.15455	.40658	-.69644	1.00553

	Equal variances not assumed			.379	18.567	.709	.15455	.40765	-.70003	1.00913
Content Quality	Equal variances assumed	.506	.486	.511	19	.615	.15909	.31153	-.49296	.81114
	Equal variances not assumed			.506	17.766	.619	.15909	.31412	-.50146	.81965
Trustworthiness of Content	Equal variances assumed	.407	.531	1.119	19	.277	.30455	.27226	-.26530	.87439
	Equal variances not assumed			1.105	17.027	.285	.30455	.27560	-.27685	.88594

Hypothesis Test Summary		
Null Hypothesis	Sig.	Decision
The distribution of Dependability is the same across categories of Treatment.	.223	Retain the null hypothesis.
The distribution of Novelty is the same across categories of Treatment.	.918	Retain the null hypothesis.

Female population – between groups

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Attractiveness	Equal variances assumed	.053	.821	.166	17	.870	.05278	.31739	-.61685	.72241
	Equal variances not assumed			.167	16.898	.870	.05278	.31679	-.61590	.72145
Efficiency	Equal variances assumed	.173	.683	.159	17	.875	.06111	.38384	-.74871	.87093
	Equal variances not assumed			.161	16.771	.874	.06111	.37903	-.73941	.86163
Dependability	Equal variances assumed	.370	.551	.658	17	.520	.28056	.42658	-.61945	1.18056
	Equal variances not assumed			.652	15.950	.523	.28056	.43007	-.63139	1.19250
Usefulness	Equal variances assumed	.085	.774	-.138	17	.892	-.07500	.54172	-1.21793	1.06793
	Equal variances not assumed			-.139	16.942	.891	-.07500	.54013	-1.21487	1.06487
Novelty	Equal variances assumed	.027	.872	-.814	17	.427	-.48889	.60051	-1.75586	.77808

	Equal variances not assumed			-.803	15.113	.434	-.48889	.60878	-1.78562	.80785
Trustworthiness of Content	Equal variances assumed	3.434	.081	-.601	17	.556	-.21667	.36067	-.97761	.54428
	Equal variances not assumed			-.586	13.219	.567	-.21667	.36950	-1.01357	.58024
Content Quality	Equal variances assumed	.517	.482	.716	17	.484	.23056	.32208	-.44898	.91009
	Equal variances not assumed			.707	15.364	.490	.23056	.32601	-.46288	.92399

Null Hypothesis	**Test**	**Sig.**	**Decision**
The distribution of Perspicuity is the same across categories of Treatment.	Independent-Samples Mann-Whitney U Test	.661[a]	Retain the null hypothesis.

Female population – Knew or Not it was made of AI

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Attractiveness	Equal variances assumed	.165	.689	.520	17	.610	.16964	.32621	-.51860	.85789
	Equal variances not assumed			.496	10.956	.630	.16964	.34222	-.58396	.92325
Efficiency	Equal variances assumed	.386	.542	.112	17	.912	.04464	.39746	-.79392	.88320
	Equal variances not assumed			.104	10.144	.919	.04464	.42778	-.90668	.99597
Perspicuity	Equal variances assumed	.353	.560	-.623	17	.542	-.27381	.43960	-1.20127	.65366
	Equal variances not assumed			-.692	16.514	.498	-.27381	.39549	-1.11009	.56247
Dependability	Equal variances assumed	.087	.772	-.180	17	.859	-.08036	.44671	-1.02283	.86212
	Equal variances not assumed			-.179	12.395	.861	-.08036	.45013	-1.05764	.89693
Usefulness	Equal variances assumed	1.992	.176	-.212	17	.834	-.11905	.56031	-1.30119	1.06309

	Equal variances not assumed			-.228	15.414	.823	-.11905	.52246	-1.23005	.99195
Novelty	Equal variances assumed	.148	.705	4.128	17	.001	1.84821	.44777	.90350	2.79293
	Equal variances not assumed			3.865	10.383	.003	1.84821	.47819	.78804	2.90839
Trustworthiness of Content	Equal variances assumed	3.886	.065	.525	17	.606	.19643	.37425	-.59317	.98602
	Equal variances not assumed			.605	16.998	.553	.19643	.32488	-.48902	.88188
Content Quality	Equal variances assumed	1.889	.187	-.362	17	.722	-.12202	.33708	-.83319	.58914
	Equal variances not assumed			-.405	16.657	.691	-.12202	.30130	-.75871	.51467

Mix population – all, by treatment

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Attractiveness	Equal variances assumed	2.667	.114	.340	27	.736	.09167	.26952	-.46133	.64467
	Equal variances not assumed			.344	25.387	.734	.09167	.26645	-.45667	.64000
Efficiency	Equal variances assumed	.450	.508	.765	27	.451	.28214	.36873	-.47443	1.03872
	Equal variances not assumed			.772	25.998	.447	.28214	.36525	-.46864	1.03292
Dependability	Equal variances assumed	.013	.909	1.294	27	.207	.43810	.33857	-.25659	1.13278
	Equal variances not assumed			1.298	26.992	.205	.43810	.33750	-.25440	1.13060
Usefulness	Equal variances assumed	.308	.583	-.058	27	.954	-.02262	.38685	-.81636	.77112

	Equal variances not assumed			-.059	26.907	.954	-.02262	.38506	-.81282	.76758
Novelty	Equal variances assumed	.389	.538	-.014	27	.989	-.00714	.50095	-1.03500	1.02071
	Equal variances not assumed			-.014	24.052	.989	-.00714	.50601	-1.05138	1.03710
Trustworthiness of Content	Equal variances assumed	1.357	.254	.556	27	.583	.15595	.28051	-.41960	.73151
	Equal variances not assumed			.550	24.081	.587	.15595	.28333	-.42870	.74060
Content Quality	Equal variances assumed	.208	.652	1.201	27	.240	.28929	.24095	-.20510	.78367
	Equal variances not assumed			1.195	25.960	.243	.28929	.24206	-.20832	.78689

Null Hypothesis	Sig.	Decision
The distribution of Perspicuity is the same across categories of Group.	.747[a]	Retain the null hypothesis.

Mix population – all, by gender

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Attractiveness	Equal variances assumed	.459	.506	-.503	19	.621	-.15682	.31207	-.81000	.49636
	Equal variances not assumed			-.496	16.975	.626	-.15682	.31598	-.82356	.50992
Efficiency	Equal variances assumed	.455	.508	-1.981	19	.062	-.82500	.41640	-1.69653	.04653
	Equal variances not assumed			-1.943	15.366	.071	-.82500	.42470	-1.72836	.07836
Usefulness	Equal variances assumed	1.140	.299	.804	19	.431	.32273	.40135	-.51732	1.16277
	Equal variances not assumed			.818	17.842	.424	.32273	.39439	-.50638	1.15183
Novelty	Equal variances	.209	.653	-.199	19	.844	-.13182	.66291	-1.51932	1.25568

	assumed									
	Equal variances not assumed			-.200	18.996	.844	-.13182	.66006	-1.51337	1.24973
Content Quality	Equal variances assumed	.018	.895	-1.325	19	.201	-.39773	.30010	-1.02584	.23039
	Equal variances not assumed			-1.319	18.329	.203	-.39773	.30148	-1.03030	.23484
Trustworthiness of Content	Equal variances assumed	.036	.851	-.065	19	.949	-.01818	.28105	-.60643	.57007
	Equal variances not assumed			-.065	19.000	.949	-.01818	.27970	-.60361	.56725

Null Hypothesis	Test	Sig.	Decision
The distribution of Perspicuity is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.863[a]	Retain the null hypothesis.
The distribution of Dependability is the same across categories of Gender.	Independent-Samples Mann-Whitney U Test	.223[a]	Retain the null hypothesis.

9.5.5. Independent Samples Effect Size – Cohen's *d*

Mix population – LCOMU2812, no treatment group

	Standardizer	Cohen's *d*	95% Confidence Interval	
			Lower	Upper
Attractiveness	.81947	-.031	-1.217	1.157
Efficiency	1.122	-.594	-1.796	.639
Perspicuity	.710	-.023	-1.210	1.164
Dependability	.988	-.439	-1.630	.777
Usefulness	.945	.282	-.919	1.468
Novelty	1.312	.381	-.829	1.570
Content Quality	.573	.145	-1.047	1.330
Trustworthiness of Content	.675	.383	-.827	1.572

Table 53 Effect Size of the mixed population – LCOMU2812, no treatment group

Mix LCOMU2812 – treatment groups

	Standardizer	Cohen's *d*	95% Confidence Interval	
			Lower	Upper
Attractiveness	.71893	.019	-.838	.875
Efficiency	1.008	.534	-.345	1.401
Perspicuity	.915	.256	-.608	1.113
Dependability	.968	.580	-.303	1.448
Usefulness	.931	.166	-.694	1.022
Novelty	1.514	.150	-.710	1.006
Content Quality	.73951	-.135	-1.372	1.110
Trustworthiness of Content	.54199	-2.030	-3.570	-.413

Mix LCOMU2812 – by treatment

	Standardizer	Cohen's *d*	95% Confidence Interval	
			Lower	Upper
Attractiveness	.71893	.019	-.838	.875
Efficiency	1.00791	.534	-.345	1.401
Perspicuity	.91523	.256	-.608	1.113
Dependability	.96795	.580	-.303	1.448
Usefulness	.93054	.166	-.694	1.022
Novelty	1.51409	.150	-.710	1.006
Content Quality	.62312	.489	-.388	1.353
Trustworthiness of Content	.71300	.223	-.639	1.079

Female population – treatment groups

	Standardizer	Cohen's *d*	95% Confidence Interval	
			Lower	Upper
Attractiveness	.69077	.076	-.826	.976
Efficiency	.83539	.073	-.829	.973
Perspicuity	.93191	.149	-.755	1.049
Dependability	.92842	.302	-.608	1.204
Usefulness	1.17901	-.064	-.963	.838
Novelty	1.30697	-.374	-1.278	.540
Trustworthiness of Content	.78497	-.276	-1.177	.633
Content Quality	.70099	.329	-.583	1.231

Female population – AI noticed

	Standardizer	Cohen's *d*	95% Confidence Interval	
			Lower	Upper
Attractiveness	.68590	.247	-.692	1.180
Efficiency	.83570	.053	-.880	.985
Perspicuity	.92431	-.296	-1.229	.645
Dependability	.93927	-.086	-1.017	.848
Usefulness	1.17812	-.101	-1.032	.833
Novelty	.94150	1.963	.805	3.083
Trustworthiness of Content	.78691	.250	-.690	1.182
Content Quality	.70875	-.172	-1.104	.764

Mix all – by treatment

	Standardizer	Cohen's *d*	95% Confidence Interval	
			Lower	Upper
Attractiveness	.72526	.126	-.604	.854
Efficiency	.99225	.284	-.450	1.014
Perspicuity	.84869	.087	-.643	.815
Dependability	.91108	.481	-.263	1.216
Usefulness	1.04100	-.022	-.750	.707
Novelty	1.34803	-.005	-.734	.723
Content Quality	.64839	.446	-.296	1.180
Trustworthiness of Content	.75484	.207	-.526	.935

Mix all – by gender

	Standardizer	Cohen's *d*	95% Confidence Interval	
			Lower	Upper
Attractiveness	.71831	-.313	-1.081	.460
Efficiency	.93319	-.799	-1.587	.002
Perspicuity	.84954	-.011	-.776	.755
Dependability	.90013	-.602	-1.379	.185
Usefulness	1.02754	.330	-.443	1.098
Novelty	1.34744	-.061	-.826	.706
Content Quality	.66542	-.024	-.789	.742
Trustworthiness of Content	.75409	-.236	-1.002	.535

9.5.6. MANOVA – all participants

Treatment x path

A. Between-Subjects Factors

	Value Label	N
Treatment	No Treatment	15
	Treatment	14
Path in App	Wrong	10
	Good	19

B. Box's Test of Equality of Covariance Matrices

Box's M	145.549
F	1.885
df1	36
df2	945.537
Sig.	.001

C. Levene's Test of Equality of Error Variances

		Levene Statistic	df1	df2	Sig.
Attractiveness	Based on Mean	1.077	3	25	.377
	Based on Median	1.047	3	25	.389
	Based on Median and with adjusted df	1.047	3	22.511	.391
	Based on trimmed mean	1.090	3	25	.372
Efficiency	Based on Mean	1.556	3	25	.225
	Based on Median	1.477	3	25	.245
	Based on Median and with adjusted df	1.477	3	19.413	.252
	Based on trimmed mean	1.563	3	25	.223
Perspicuity	Based on Mean	.986	3	25	.415
	Based on Median	.573	3	25	.638
	Based on Median and with adjusted df	.573	3	17.187	.641
	Based on trimmed mean	.986	3	25	.415
Dependability	Based on Mean	1.704	3	25	.192
	Based on Median	1.667	3	25	.199
	Based on Median and with adjusted df	1.667	3	21.348	.204
	Based on trimmed mean	1.762	3	25	.180
Usefulness	Based on Mean	.154	3	25	.926
	Based on Median	.146	3	25	.931
	Based on Median and with adjusted df	.146	3	24.870	.931
	Based on trimmed mean	.145	3	25	.932
Novelty	Based on Mean	.864	3	25	.473
	Based on Median	.433	3	25	.731
	Based on Median and with adjusted df	.433	3	21.420	.732
	Based on trimmed mean	.798	3	25	.507
Content Quality	Based on Mean	.239	3	25	.868
	Based on Median	.208	3	25	.890
	Based on Median and with adjusted df	.208	3	23.317	.890

	Based on trimmed mean	.238	3	25	.869
Trustworthiness of Content	Based on Mean	2.296	3	25	.102
	Based on Median	1.178	3	25	.338
	Based on Median and with adjusted df	1.178	3	18.841	.345
	Based on trimmed mean	2.336	3	25	.098

D. Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Pillai's Trace	.311	1.016[b]	8.000	18.000	.459	.311	8.129	.331
	Wilks' Lambda	.689	1.016[b]	8.000	18.000	.459	.311	8.129	.331
	Hotelling's Trace	.452	1.016[b]	8.000	18.000	.459	.311	8.129	.331
	Roy's Largest Root	.452	1.016[b]	8.000	18.000	.459	.311	8.129	.331
Path in App	Pillai's Trace	.348	1.202[b]	8.000	18.000	.352	.348	9.613	.392
	Wilks' Lambda	.652	1.202[b]	8.000	18.000	.352	.348	9.613	.392
	Hotelling's Trace	.534	1.202[b]	8.000	18.000	.352	.348	9.613	.392
	Roy's Largest Root	.534	1.202[b]	8.000	18.000	.352	.348	9.613	.392
Treatment * PathApp	Pillai's Trace	.349	1.204[b]	8.000	18.000	.350	.349	9.632	.392
	Wilks' Lambda	.651	1.204[b]	8.000	18.000	.350	.349	9.632	.392
	Hotelling's Trace	.535	1.204[b]	8.000	18.000	.350	.349	9.632	.392
	Roy's Largest Root	.535	1.204[b]	8.000	18.000	.350	.349	9.632	.392

E. Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Attractiveness	.262	1	.262	.488	.491	.019	.488	.103
	Efficiency	1.823	1	1.823	2.138	.156	.079	2.138	.290
	Perspicuity	.062	1	.062	.083	.775	.003	.083	.059
	Dependability	1.814	1	1.814	2.252	.146	.083	2.252	.303
	Usefulness	.000	1	.000	.000	.987	.000	.000	.050
	Novelty	.240	1	.240	.133	.719	.005	.133	.064
	Content Quality	.628	1	.628	1.473	.236	.056	1.473	.215
	Trustworthiness of Content	.427	1	.427	.760	.392	.029	.760	.134
PathApp	Attractiveness	.000	1	.000	.000	1.000	.000	.000	.050
	Efficiency	.907	1	.907	1.063	.312	.041	1.063	.168
	Perspicuity	.815	1	.815	1.097	.305	.042	1.097	.172
	Dependability	1.627	1	1.627	2.020	.168	.075	2.020	.277
	Usefulness	.007	1	.007	.006	.940	.000	.006	.051
	Novelty	1.277	1	1.277	.705	.409	.027	.705	.127
	Content Quality	.651	1	.651	1.527	.228	.058	1.527	.221
	Trustworthiness of Content	.567	1	.567	1.009	.325	.039	1.009	.162
Treatment *	Attractiveness	.802	1	.802	1.496	.233	.056	1.496	.218

PathApp	Efficiency	4.425	1	4.425	5.188	**.032**	.172	5.188	.591
	Perspicuity	.069	1	.069	.093	.763	.004	.093	.060
	Dependability	.685	1	.685	.851	.365	.033	.851	.144
	Usefulness	.057	1	.057	.049	.827	.002	.049	.055
	Novelty	2.422	1	2.422	1.337	.259	.051	1.337	.199
	Content Quality	.053	1	.053	.125	.726	.005	.125	.063
	Trustworthiness of Content	.777	1	.777	1.380	.251	.052	1.380	.204

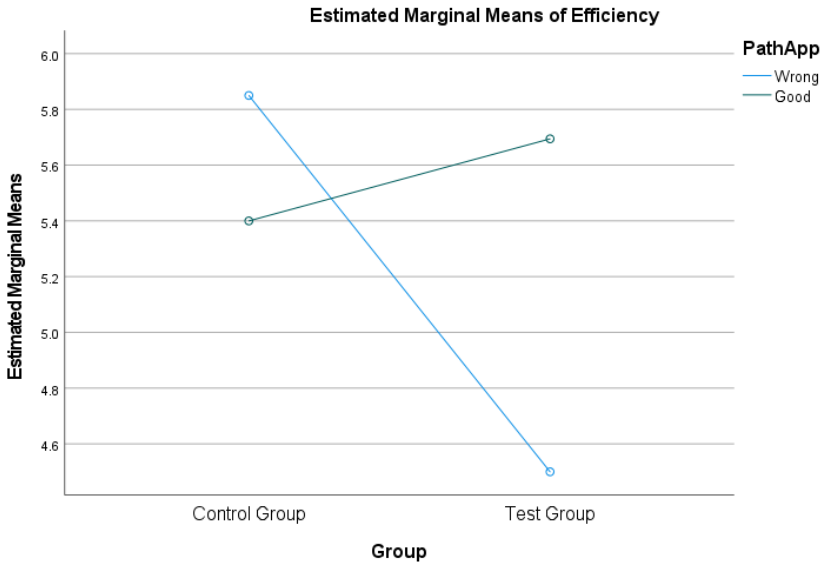


Table 54 Treatment x Path in App for Efficiency

Treatment x covid check

A. Between-Subjects Factors

	Value Label	N
Treatment	No Treatment	15
	Treatment	14
Covid Check	No covid check	24
	Covid check	5

B. Box's Test of Equality of Covariance Matrices

Box's M	59.599
F	.988
df1	36
df2	1628.589
Sig.	.490

C. Levene's Test of Equality of Error Variances

		Levene Statistic	df1	df2	Sig.
Attractiveness	Based on Mean	1.910	3	25	.154
	Based on Median	1.647	3	25	.204
	Based on Median and with adjusted df	1.647	3	18.607	.213
	Based on trimmed mean	1.901	3	25	.155
Efficiency	Based on Mean	5.011	3	25	.007
	Based on Median	1.043	3	25	.391
	Based on Median and with adjusted df	1.043	3	5.198	.447
	Based on trimmed mean	4.696	3	25	.010
Perspicuity	Based on Mean	.942	3	25	.435
	Based on Median	.311	3	25	.817
	Based on Median and with adjusted df	.311	3	14.932	.817
	Based on trimmed mean	.798	3	25	.507
Dependability	Based on Mean	2.543	3	25	.079
	Based on Median	.387	3	25	.763
	Based on Median and with adjusted df	.387	3	10.779	.765
	Based on trimmed mean	2.314	3	25	.100
Usefulness	Based on Mean	.967	3	25	.424
	Based on Median	.593	3	25	.625
	Based on Median and with adjusted df	.593	3	18.646	.627
	Based on trimmed mean	.946	3	25	.433
Novelty	Based on Mean	.414	3	25	.744
	Based on Median	.457	3	25	.715
	Based on Median	.457	3	19.1	.715

	and with adjusted df			16	
	Based on trimmed mean	.386	3	25	.764
Content Quality	Based on Mean	2.339	3	25	.098
	Based on Median	2.451	3	25	.087
	Based on Median and with adjusted df	2.451	3	20.699	.092
	Based on trimmed mean	2.348	3	25	.097
Trustworthiness of Content	Based on Mean	2.460	3	25	.086
	Based on Median	1.585	3	25	.218
	Based on Median and with adjusted df	1.585	3	19.481	.225
	Based on trimmed mean	2.397	3	25	.092

D. Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Pillai's Trace	.175	.477[b]	8.000	18.000	.856	.175	3.818	.163
	Wilks' Lambda	.825	.477[b]	8.000	18.000	.856	.175	3.818	.163
	Hotelling's Trace	.212	.477[b]	8.000	18.000	.856	.175	3.818	.163
	Roy's Largest Root	.212	.477[b]	8.000	18.000	.856	.175	3.818	.163
Covid Check	Pillai's Trace	.364	1.286[b]	8.000	18.000	.311	.364	10.290	.419
	Wilks' Lambda	.636	1.286[b]	8.000	18.000	.311	.364	10.290	.419
	Hotelling's Trace	.572	1.286[b]	8.000	18.000	.311	.364	10.290	.419
	Roy's Largest Root	.572	1.286[b]	8.000	18.000	.311	.364	10.290	.419
Treatment * Covid Check	Pillai's Trace	.320	1.060[b]	8.000	18.000	.431	.320	8.480	.345
	Wilks' Lambda	.680	1.060[b]	8.000	18.000	.431	.320	8.480	.345
	Hotelling's Trace	.471	1.060[b]	8.000	18.000	.431	.320	8.480	.345
	Roy's Largest Root	.471	1.060[b]	8.000	18.000	.431	.320	8.480	.345

E. Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Attractiveness	.766	1	.766	1.459	.238	.055	1.459	.213
	Efficiency	.594	1	.594	.561	.461	.022	.561	.111
	Perspicuity	1.174	1	1.174	1.728	.201	.065	1.728	.244
	Dependability	1.460	1	1.460	1.670	.208	.063	1.670	.237
	Usefulness	2.573	1	2.573	2.784	.108	.100	2.784	.361
	Novelty	.532	1	.532	.283	.599	.011	.283	.081
	Content Quality	1.562	1	1.562	4.171	.052	.143	4.171	.502
	Trustworthiness of Content	.879	1	.879	1.548	.225	.058	1.548	.223
Covid Check	Attractiveness	.016	1	.016	.030	.864	.001	.030	.053
	Efficiency	.000	1	.000	.000	.984	.000	.000	.050
	Perspicuity	1.085	1	1.085	1.598	.218	.060	1.598	.229
	Dependability	.174	1	.174	.199	.660	.008	.199	.071
	Usefulness	.125	1	.125	.136	.716	.005	.136	.065
	Novelty	.417	1	.417	.222	.641	.009	.222	.074
	Content Quality	.502	1	.502	1.339	.258	.051	1.339	.200
	Trustworthiness of Content	.098	1	.098	.172	.682	.007	.172	.068

Treatment * Covid Check	Attractiveness	1.085	1	1.085	2.068	.163	.076	2.068	.282
	Efficiency	.098	1	.098	.092	.764	.004	.092	.060
	Perspicuity	1.778	1	1.778	2.618	.118	.095	2.618	.343
	Dependability	.293	1	.293	.336	.568	.013	.336	.086
	Usefulness	6.146	1	6.146	6.649	**.016**	.210	6.649	.698
	Novelty	1.410	1	1.410	.751	.394	.029	.751	.133
	Content Quality	1.174	1	1.174	3.133	.089	.111	3.133	.398
	Trustworthiness of Content	.959	1	.959	1.689	.206	.063	1.689	.240

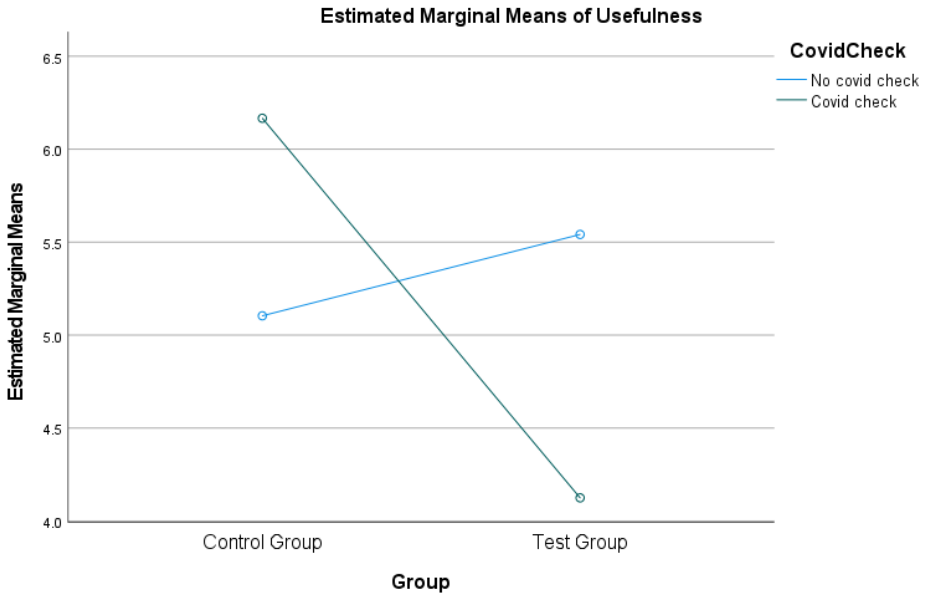


Table 55 Treatment x Covid check for Usefulness

Treatment x gender

A. Between-Subjects Factors

	Value Label	N
Treatment	No Treatment Treatment	15
	Treatment Treatment	14
Gender	Male	10
	Female	19

B. Box's Test of Equality of Covariance Matrices

Box's M	115.643
F	1.498
df1	36
df2	945.537
Sig.	.031

C. Levene's Test of Equality of Error Variances

		Levene Statistic	df1	df2	Sig.
Attractiveness	Based on Mean	2.267	3	25	.105
	Based on Median	.702	3	25	.560
	Based on Median and with adjusted df	.702	3	13.500	.567
	Based on trimmed mean	2.154	3	25	.119
Efficiency	Based on Mean	1.377	3	25	.273
	Based on Median	1.072	3	25	.379
	Based on Median and with adjusted df	1.072	3	17.574	.386
	Based on trimmed mean	1.398	3	25	.267
Perspicuity	Based on Mean	1.507	3	25	.237
	Based on Median	.470	3	25	.706
	Based on Median and with adjusted df	.470	3	15.295	.707
	Based on trimmed mean	1.392	3	25	.268
Dependability	Based on Mean	2.745	3	25	.064
	Based on Median	2.210	3	25	.112
	Based on Median and with adjusted df	2.210	3	19.013	.120
	Based on trimmed mean	2.805	3	25	.060
Usefulness	Based on Mean	.880	3	25	.465
	Based on Median	.947	3	25	.433
	Based on Median and with adjusted df	.947	3	23.893	.433
	Based on trimmed mean	.886	3	25	.462
Novelty	Based on Mean	.204	3	25	.893
	Based on Median	.111	3	25	.953
	Based on Median and with adjusted	.111	3	20.557	.953

	df				
	Based on trimmed mean	.213	3	25	.886
Content Quality	Based on Mean	.365	3	25	.779
	Based on Median	.365	3	25	.779
	Based on Median and with adjusted df	.365	3	23.934	.779
	Based on trimmed mean	.357	3	25	.785
Trustworthiness of Content	Based on Mean	1.873	3	25	.160
	Based on Median	1.482	3	25	.243
	Based on Median and with adjusted df	1.482	3	22.811	.246
	Based on trimmed mean	1.922	3	25	.152

D. Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Pillai's Trace	.263	.802	8.000	18.000	.609	.263	6.418	.262
	Wilks' Lambda	.737	.802	8.000	18.000	.609	.263	6.418	.262
	Hotelling's Trace	.357	.802	8.000	18.000	.609	.263	6.418	.262
	Roy's Largest Root	.357	.802	8.000	18.000	.609	.263	6.418	.262
Gender	Pillai's Trace	.471	2.006	8.000	18.000	.105	.471	16.049	.633
	Wilks' Lambda	.529	2.006	8.000	18.000	.105	.471	16.049	.633
	Hotelling's Trace	.892	2.006	8.000	18.000	.105	.471	16.049	.633
	Roy's Largest Root	.892	2.006	8.000	18.000	.105	.471	16.049	.633
Treatment * Gender	Pillai's Trace	.206	.585	8.000	18.000	.777	.206	4.682	.194
	Wilks' Lambda	.794	.585	8.000	18.000	.777	.206	4.682	.194
	Hotelling's Trace	.260	.585	8.000	18.000	.777	.206	4.682	.194
	Roy's Largest Root	.260	.585	8.000	18.000	.777	.206	4.682	.194

E. Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Attractiveness	.067	1	.067	.121	.730	.005	.121	.063
	Efficiency	.827	1	.827	.922	.346	.036	.922	.152
	Perspicuity	.013	1	.013	.017	.898	.001	.017	.052
	Dependability	1.573	1	1.573	1.940	.176	.072	1.940	.268
	Usefulness	.001	1	.001	.001	.976	.000	.001	.050
	Novelty	.277	1	.277	.151	.701	.006	.151	.066
	Content Quality	.651	1	.651	1.439	.242	.054	1.439	.211
	Trustworthiness of Content	.656	1	.656	1.231	.278	.047	1.231	.187
Gender	Attractiveness	.327	1	.327	.590	.450	.023	.590	.115
	Efficiency	3.627	1	3.627	4.042	.055	.139	4.042	.489
	Perspicuity	.000	1	.000	.000	.987	.000	.000	.050
	Dependability	1.872	1	1.872	2.307	.141	.084	2.307	.309
	Usefulness	.746	1	.746	.655	.426	.026	.655	.122
	Novelty	.058	1	.058	.032	.860	.001	.032	.053
	Content Quality	.001	1	.001	.001	.971	.000	.001	.050
	Trustworthiness of Content	.220	1	.220	.413	.526	.016	.413	.095
Treatment *	Attractiveness	.015	1	.015	.028	.869	.001	.028	.053

Gender	Efficiency	.567	1	.567	.632	.434	.025	.632	.119
	Perspicuity	.058	1	.058	.075	.786	.003	.075	.058
	Dependability	.288	1	.288	.355	.557	.014	.355	.088
	Usefulness	.050	1	.050	.044	.836	.002	.044	.055
	Novelty	3.157	1	3.157	1.721	.202	.064	1.721	.243
	Content Quality	.047	1	.047	.104	.750	.004	.104	.061
	Trustworthiness of Content	1.862	1	1.862	3.493	.073	.123	3.493	.435

Treatment x Related interest

A. Between-Subjects Factors

	Value Label	N
Treatment	No Treatment	15
	Treatment	14
Related Interest	Has no related interest	13
	Has at least one related interest	16

B. Levene's Test of Equality of Error Variances

		Levene Statistic	df1	df2	Sig.
Attractiveness	Based on Mean	.620	3	25	.609
	Based on Median	.455	3	25	.716
	Based on Median and with adjusted df	.455	3	17.885	.717
	Based on trimmed mean	.605	3	25	.618
Efficiency	Based on Mean	.750	3	25	.533
	Based on Median	.521	3	25	.672
	Based on Median and with adjusted df	.521	3	17.770	.673
	Based on trimmed mean	.722	3	25	.548
Perspicuity	Based on Mean	1.717	3	25	.189
	Based on Median	1.733	3	25	.186
	Based on Median and with adjusted df	1.733	3	16.455	.199
	Based on trimmed mean	1.817	3	25	.170
Dependability	Based on Mean	1.478	3	25	.245
	Based on Median	1.462	3	25	.249
	Based on Median and with adjusted df	1.462	3	21.254	.253
	Based on trimmed mean	1.498	3	25	.239
Usefulness	Based on Mean	1.829	3	25	.168
	Based on Median	1.776	3	25	.177
	Based on Median and with adjusted df	1.776	3	24.773	.178
	Based on trimmed mean	1.836	3	25	.166
Novelty	Based on Mean	.622	3	25	.608
	Based on Median	.345	3	25	.793
	Based on Median and with adjusted df	.345	3	16.819	.793
	Based on trimmed mean	.536	3	25	.662
Content Quality	Based on Mean	.895	3	25	.458
	Based on Median	.791	3	25	.510
	Based on Median	.791	3	21.3	.512

	and with adjusted df			37	
	Based on trimmed mean	.881	3	25	.464
Trustworthiness of Content	Based on Mean	3.753	3	25	.024
	Based on Median	3.479	3	25	.031
	Based on Median and with adjusted df	3.479	3	20.791	.034
	Based on trimmed mean	3.749	3	25	.024

C. Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Pillai's Trace	.200	.562[b]	8.000	18.000	.795	.200	4.493	.187
	Wilks' Lambda	.800	.562[b]	8.000	18.000	.795	.200	4.493	.187
	Hotelling's Trace	.250	.562[b]	8.000	18.000	.795	.200	4.493	.187
	Roy's Largest Root	.250	.562[b]	8.000	18.000	.795	.200	4.493	.187
Related Interest	Pillai's Trace	.166	.449[b]	8.000	18.000	.876	.166	3.592	.155
	Wilks' Lambda	.834	.449[b]	8.000	18.000	.876	.166	3.592	.155
	Hotelling's Trace	.200	.449[b]	8.000	18.000	.876	.166	3.592	.155
	Roy's Largest Root	.200	.449[b]	8.000	18.000	.876	.166	3.592	.155
Treatment * Related Interest	Pillai's Trace	.425	1.662[b]	8.000	18.000	.176	.425	13.294	.536
	Wilks' Lambda	.575	1.662[b]	8.000	18.000	.176	.425	13.294	.536
	Hotelling's Trace	.739	1.662[b]	8.000	18.000	.176	.425	13.294	.536
	Roy's Largest Root	.739	1.662[b]	8.000	18.000	.176	.425	13.294	.536

D. Tests of Between-Subjects Effect

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Attractiveness	.006	1	.006	.014	.908	.001	.014	.051
	Efficiency	.310	1	.310	.334	.568	.013	.334	.086
	Perspicuity	6.332E-5	1	6.332E-5	.000	.992	.000	.000	.050
	Dependability	1.058	1	1.058	1.280	.269	.049	1.280	.193
	Usefulness	.019	1	.019	.018	.896	.001	.018	.052
	Novelty	.013	1	.013	.007	.934	.000	.007	.051
	Content Quality	.496	1	.496	1.141	.296	.044	1.141	.177
	Trustworthiness of Content	.109	1	.109	.190	.667	.008	.190	.070
Related Interest	Attractiveness	.001	1	.001	.002	.963	.000	.002	.050
	Efficiency	1.434	1	1.434	1.546	.225	.058	1.546	.223
	Perspicuity	.596	1	.596	.923	.346	.036	.923	.152
	Dependability	.695	1	.695	.841	.368	.033	.841	.143
	Usefulness	1.502	1	1.502	1.356	.255	.051	1.356	.201
	Novelty	.008	1	.008	.004	.950	.000	.004	.050
	Content Quality	.289	1	.289	.664	.423	.026	.664	.123
	Trustworthiness of Content	.787	1	.787	1.369	.253	.052	1.369	.203
Treatment *	Attractiveness	2.533	1	2.533	5.431	**.028**	.178	5.431	.610

Related Interest	Efficiency	2.098	1	2.098	2.261	.145	.083	2.261	.304
	Perspicuity	2.819	1	2.819	4.368	**.047**	.149	4.368	.520
	Dependability	1.124	1	1.124	1.360	.255	.052	1.360	.202
	Usefulness	.094	1	.094	.085	.773	.003	.085	.059
	Novelty	.904	1	.904	.470	.499	.018	.470	.101
	Content Quality	.217	1	.217	.499	.487	.020	.499	.104
	Trustworthiness of Content	.255	1	.255	.444	.511	.017	.444	.098

Treatment x English level

A. Between-Subjects Factors

	Value Label	N
Treatment	No Treatment	15
	Treatment	14
English Level	Sufficient	13
	Good	16

B. Box's Test of Equality of Covariance Matrices

Box's M	81.210
F	1.052
df1	36
df2	945.537
Sig.	.388

C. Levene's Test of Equality of Error Variances

		Levene Statistic	df1	df2	Sig.
Attractiveness	Based on Mean	.660	3	25	.584
	Based on Median	.299	3	25	.826
	Based on Median and with adjusted df	.299	3	19.293	.826
	Based on trimmed mean	.609	3	25	.616
Efficiency	Based on Mean	.406	3	25	.750
	Based on Median	.438	3	25	.728
	Based on Median and with adjusted df	.438	3	18.174	.729
	Based on trimmed mean	.471	3	25	.706
Perspicuity	Based on Mean	2.509	3	25	.082
	Based on Median	2.215	3	25	.111
	Based on Median and with adjusted df	2.215	3	18.930	.120
	Based on trimmed mean	2.487	3	25	.084
Dependability	Based on Mean	1.527	3	25	.232
	Based on Median	1.356	3	25	.279
	Based on Median and with adjusted df	1.356	3	20.034	.285
	Based on trimmed mean	1.544	3	25	.228
Usefulness	Based on Mean	.452	3	25	.718
	Based on Median	.267	3	25	.849
	Based on Median and with adjusted df	.267	3	22.167	.848
	Based on trimmed mean	.469	3	25	.706
Novelty	Based on Mean	.942	3	25	.435
	Based on Median	.442	3	25	.725
	Based on Median and with adjusted	.442	3	21.911	.725

	df				
	Based on trimmed mean	.897	3	25	.456
Content Quality	Based on Mean	.855	3	25	.477
	Based on Median	.489	3	25	.693
	Based on Median and with adjusted df	.489	3	20.815	.694
	Based on trimmed mean	.846	3	25	.482
Trustworthiness of Content	Based on Mean	1.340	3	25	.284
	Based on Median	.511	3	25	.678
	Based on Median and with adjusted df	.511	3	19.299	.679
	Based on trimmed mean	1.403	3	25	.265

D. Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Pillai's Trace	.312	1.019[b]	8.000	18.000	.457	.312	8.149	.332
	Wilks' Lambda	.688	1.019[b]	8.000	18.000	.457	.312	8.149	.332
	Hotelling's Trace	.453	1.019[b]	8.000	18.000	.457	.312	8.149	.332
	Roy's Largest Root	.453	1.019[b]	8.000	18.000	.457	.312	8.149	.332
English Level	Pillai's Trace	.288	.909[b]	8.000	18.000	.530	.288	7.273	.296
	Wilks' Lambda	.712	.909[b]	8.000	18.000	.530	.288	7.273	.296
	Hotelling's Trace	.404	.909[b]	8.000	18.000	.530	.288	7.273	.296
	Roy's Largest Root	.404	.909[b]	8.000	18.000	.530	.288	7.273	.296
Treatment * English Level	Pillai's Trace	.142	.371[b]	8.000	18.000	.923	.142	2.968	.134
	Wilks' Lambda	.858	.371[b]	8.000	18.000	.923	.142	2.968	.134
	Hotelling's Trace	.165	.371[b]	8.000	18.000	.923	.142	2.968	.134
	Roy's Largest Root	.165	.371[b]	8.000	18.000	.923	.142	2.968	.134

E. Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Attractiveness	.209	1	.209	.385	.541	.015	.385	.092
	Efficiency	1.884	1	1.884	2.078	.162	.077	2.078	.284
	Perspicuity	.277	1	.277	.384	.541	.015	.384	.092
	Dependability	2.445	1	2.445	2.965	.097	.106	2.965	.381
	Usefulness	.011	1	.011	.010	.921	.000	.010	.051
	Novelty	.001	1	.001	.001	.982	.000	.001	.050
	Content Quality	.936	1	.936	2.154	.155	.079	2.154	.292
	Trustworthiness of Content	.524	1	.524	.930	.344	.036	.930	.153
English Level	Attractiveness	.566	1	.566	1.039	.318	.040	1.039	.165
	Efficiency	2.641	1	2.641	2.912	.100	.104	2.912	.375
	Perspicuity	1.338	1	1.338	1.860	.185	.069	1.860	.259
	Dependability	1.371	1	1.371	1.663	.209	.062	1.663	.237
	Usefulness	.068	1	.068	.059	.809	.002	.059	.056
	Novelty	.081	1	.081	.042	.840	.002	.042	.054
	Content Quality	.372	1	.372	.856	.364	.033	.856	.145
	Trustworthiness of	.484	1	.484	.860	.363	.033	.860	.145

	Content								
Treatment * English Level	Attractiveness	.008	1	.008	.015	.905	.001	.015	.052
	Efficiency	1.715	1	1.715	1.891	.181	.070	1.891	.262
	Perspicuity	.044	1	.044	.061	.806	.002	.061	.057
	Dependability	.617	1	.617	.748	.395	.029	.748	.132
	Usefulness	.514	1	.514	.448	.510	.018	.448	.099
	Novelty	.439	1	.439	.226	.638	.009	.226	.074
	Content Quality	.160	1	.160	.367	.550	.014	.367	.090
	Trustworthiness of Content	.953	1	.953	1.692	.205	.063	1.692	.240

Treatment x Measured Technophobia

A. Between-Subjects Factors

	Value Label	N
Treatment	No Treatment Treatment	15
	Treatment Treatment	14
Measured Technophobia	Moderaty Technophobic	3
	Midly Technophobic	12
	Not Technophobic	14

B. Levene's Test of Equality of Error Variances

		Levene Statistic	df1	df2	Sig.
Attractiveness	Based on Mean	.314	4	23	.865
	Based on Median	.202	4	23	.935
	Based on Median and with adjusted df	.202	4	15.920	.934
	Based on trimmed mean	.287	4	23	.883
Efficiency	Based on Mean	.614	4	23	.657
	Based on Median	.542	4	23	.706
	Based on Median and with adjusted df	.542	4	17.889	.707
	Based on trimmed mean	.653	4	23	.630
Perspicuity	Based on Mean	2.377	4	23	.082
	Based on Median	1.460	4	23	.247
	Based on Median and with adjusted df	1.460	4	15.326	.262
	Based on trimmed mean	2.225	4	23	.098
Dependability	Based on Mean	.265	4	23	.897
	Based on Median	.148	4	23	.962
	Based on Median and with adjusted df	.148	4	20.635	.962
	Based on trimmed mean	.238	4	23	.914
Usefulness	Based on Mean	1.756	4	23	.172
	Based on Median	.721	4	23	.587
	Based on Median and with adjusted df	.721	4	16.680	.590
	Based on trimmed mean	1.729	4	23	.178
Novelty	Based on Mean	.341	4	23	.847
	Based on Median	.209	4	23	.931
	Based on Median and with adjusted	.209	4	16.736	.930

	df				
	Based on trimmed mean	.309	4	23	.869
Content Quality	Based on Mean	.963	4	23	.446
	Based on Median	.846	4	23	.510
	Based on Median and with adjusted df	.846	4	12.680	.521
	Based on trimmed mean	.958	4	23	.449
Trustworthiness of Content	Based on Mean	1.313	4	23	.295
	Based on Median	.470	4	23	.757
	Based on Median and with adjusted df	.470	4	16.796	.757
	Based on trimmed mean	1.323	4	23	.291

C. Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Pillai's Trace	.136	.316[b]	8.000	16.000	.948	.136	2.529	.116
	Wilks' Lambda	.864	.316[b]	8.000	16.000	.948	.136	2.529	.116
	Hotelling's Trace	.158	.316[b]	8.000	16.000	.948	.136	2.529	.116
	Roy's Largest Root	.158	.316[b]	8.000	16.000	.948	.136	2.529	.116
Measured Technophobia	Pillai's Trace	.730	1.222	16.000	34.000	.302	.365	19.549	.623
	Wilks' Lambda	.384	1.226[b]	16.000	32.000	.302	.380	19.622	.615
	Hotelling's Trace	1.305	1.223	16.000	30.000	.307	.395	19.569	.602
	Roy's Largest Root	1.010	2.146[c]	8.000	17.000	.088	.502	17.166	.657
Treatment * Measured Technophobia	Pillai's Trace	.878	1.662	16.000	34.000	.105	.439	26.598	.792
	Wilks' Lambda	.297	1.672[b]	16.000	32.000	.106	.455	26.748	.785
	Hotelling's Trace	1.782	1.671	16.000	30.000	.110	.471	26.731	.774
	Roy's Largest Root	1.345	2.857[c]	8.000	17.000	.033	.573	22.857	.802

D. Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Attractiveness	.003	1	.003	.006	.941	.000	.006	.051
	Efficiency	.263	1	.263	.230	.636	.010	.230	.075
	Perspicuity	1.353	1	1.353	2.538	.125	.099	2.538	.333
	Dependability	.242	1	.242	.256	.618	.011	.256	.077
	Usefulness	.404	1	.404	.369	.550	.016	.369	.090
	Novelty	.027	1	.027	.014	.908	.001	.014	.051
	Content Quality	.200	1	.200	.702	.411	.030	.702	.127
	Trustworthiness of Content	1.830E-5	1	1.830E-5	.000	.995	.000	.000	.050
Measured Technophobia	Attractiveness	2.895	2	1.448	3.203	.059	.218	6.407	.554
	Efficiency	.041	2	.021	.018	.982	.002	.036	.052
	Perspicuity	6.834	2	3.417	6.407	**.006**	.358	12.814	.860
	Dependability	.029	2	.015	.016	.985	.001	.031	.052
	Usefulness	2.131	2	1.066	.973	.393	.078	1.946	.198
	Novelty	2.845	2	1.422	.712	.501	.058	1.424	.155
	Content Quality	1.949	2	.974	3.428	**.050**	.230	6.856	.585
	Trustworthiness of Content	1.997	2	.998	1.827	.183	.137	3.654	.341
Treatment *	Attractiveness	.373	2	.187	.413	.666	.035	.827	.109

Measured Technophobia	Efficiency	.193	2	.096	.084	.920	.007	.168	.061
	Perspicuity	1.186	2	.593	1.112	.346	.088	2.223	.221
	Dependability	.535	2	.268	.284	.756	.024	.567	.089
	Usefulness	2.158	2	1.079	.985	.389	.079	1.970	.200
	Novelty	.513	2	.256	.128	.880	.011	.257	.067
	Content Quality	2.753	2	1.376	4.842	**.018**	.296	9.684	.744
	Trustworthiness of Content	1.228	2	.614	1.123	.342	.089	2.247	.223

Treatment x Self-reported Technophobia

A. Between-Subjects Factors

	Value Label	N
Treatment	No Treatment	15
	Treatment	14
If 'technophobia' is defined as feeling discomfort about computers or any new technology, which of the following best describes you?	Highly Technophobic	2
	Moderately Technophobic	1
	Midly Technophobic	4
	Not Technophobic	22

B. Box's Test of Equality of Covariance Matrices

Box's M	66.396
F	1.007
df1	36
df2	1242.623
Sig.	.459

C. Levene's Test of Equality of Error Variances

		Levene Statistic	df1	df2	Sig.
Attractiveness	Based on Mean	1.806	4	23	.162
	Based on Median	1.763	4	23	.171
	Based on Median and with adjusted df	1.763	4	18.891	.178
	Based on trimmed mean	1.805	4	23	.162
Efficiency	Based on Mean	.352	4	23	.840
	Based on Median	.320	4	23	.862
	Based on Median and with adjusted df	.320	4	18.460	.861
	Based on trimmed mean	.354	4	23	.839
Perspicuity	Based on Mean	2.740	4	23	.053
	Based on Median	.640	4	23	.640
	Based on Median and with adjusted df	.640	4	19.498	.641
	Based on trimmed mean	2.285	4	23	.091
Dependability	Based on Mean	1.000	4	23	.428
	Based on Median	.415	4	23	.796
	Based on Median and with adjusted df	.415	4	19.288	.796
	Based on trimmed mean	.965	4	23	.446
Usefulness	Based on Mean	1.684	4	23	.188
	Based on Median	1.545	4	23	.222
	Based on Median and with adjusted df	1.545	4	19.599	.228
	Based on trimmed mean	1.687	4	23	.187
Novelty	Based on Mean	2.227	4	23	.098
	Based on Median	1.958	4	23	.135
	Based on Median and with adjusted	1.958	4	18.986	.142

	df				
	Based on trimmed mean	2.215	4	23	.099
Content Quality	Based on Mean	.984	4	23	.436
	Based on Median	.907	4	23	.476
	Based on Median and with adjusted df	.907	4	19.823	.479
	Based on trimmed mean	.979	4	23	.438
Trustworthiness of Content	Based on Mean	.360	4	23	.835
	Based on Median	.212	4	23	.929
	Based on Median and with adjusted df	.212	4	18.532	.929
	Based on trimmed mean	.350	4	23	.841

D. Multivariate Tests

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Pillai's Trace	.464	1.731[b]	8.000	16.000	.167	.464	13.849	.535
	Wilks' Lambda	.536	1.731[b]	8.000	16.000	.167	.464	13.849	.535
	Hotelling's Trace	.866	1.731[b]	8.000	16.000	.167	.464	13.849	.535
	Roy's Largest Root	.866	1.731[b]	8.000	16.000	.167	.464	13.849	.535
Technophobia	Pillai's Trace	1.345	1.829	24.000	54.000	**.034**	.448	43.908	.949
	Wilks' Lambda	.149	1.818	24.000	47.006	.040	.470	41.677	.923
	Hotelling's Trace	2.911	1.779	24.000	44.000	.048	.492	42.688	.925
	Roy's Largest Root	1.794	4.036[c]	8.000	18.000	.007	.642	32.287	.937
Treatment * Technophobia	Pillai's Trace	.539	2.339[b]	8.000	16.000	.070	.539	18.709	.690
	Wilks' Lambda	.461	2.339[b]	8.000	16.000	.070	.539	18.709	.690
	Hotelling's Trace	1.169	2.339[b]	8.000	16.000	.070	.539	18.709	.690
	Roy's Largest Root	1.169	2.339[b]	8.000	16.000	.070	.539	18.709	.690

E. Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Treatment	Attractiveness	.275	1	.275	.524	.477	.022	.524	.107
	Efficiency	1.410	1	1.410	1.459	.239	.060	1.459	.212
	Perspicuity	3.200	1	3.200	9.162	**.006**	.285	9.162	.826
	Dependability	2.802	1	2.802	3.883	.061	.144	3.883	.471
	Usefulness	.866	1	.866	.910	.350	.038	.910	.150
	Novelty	.313	1	.313	.164	.690	.007	.164	.067
	Content Quality	1.217	1	1.217	3.003	.096	.116	3.003	.383
	Trustworthiness of Content	.017	1	.017	.028	.869	.001	.028	.053
Technophobia	Attractiveness	1.595	3	.532	1.011	.406	.117	3.034	.239
	Efficiency	1.910	3	.637	.659	.586	.079	1.976	.167
	Perspicuity	9.106	3	3.035	8.691	**.000**	.531	26.074	.985
	Dependability	5.026	3	1.675	2.322	.102	.232	6.966	.510
	Usefulness	4.340	3	1.447	1.520	.236	.165	4.559	.347
	Novelty	3.866	3	1.289	.674	.577	.081	2.022	.170
	Content Quality	1.606	3	.535	1.321	.292	.147	3.964	.305
	Trustworthiness of Content	1.183	3	.394	.649	.591	.078	1.948	.165
Treatment * Technophobia	Attractiveness	.390	1	.390	.741	.398	.031	.741	.131
	Efficiency	1.797	1	1.797	1.860	.186	.075	1.860	.257

	Perspicuity	3.566	1	3.566	10.211	**.004**	.307	10.211	.864
	Dependability	1.175	1	1.175	1.629	.215	.066	1.629	.231
	Usefulness	2.551	1	2.551	2.680	.115	.104	2.680	.348
	Novelty	1.101	1	1.101	.576	.456	.024	.576	.113
	Content Quality	.541	1	.541	1.335	.260	.055	1.335	.198
	Trustworthiness of Content	.130	1	.130	.213	.648	.009	.213	.073

9.5.7. Interview results - table overview

N	English level	Covid check	Path	Most important	Least important	Liked	Disliked	Similar app
1	Sufficient	No	Bad	Trustworthiness, Content Quality and Dependability	Attractiveness	Simple and supportive interface	Length	No
2	Sufficient	No	Bad	Trustworthiness	Not mentioned	Simple and efficient interface	Not sure if it is trustworthy	Doctissimo
3	Good	Yes	Good	Trustworthiness, confidentiality, efficiency	Novelty	Clarity (interface)	None	No
4	Sufficient	No	Good	Trustworthiness, Content Quality, Efficiency, Dependability	Novelty, Attractiveness	Contextual utility (pandemic)	Result page	Akinator
5	Good	No	Good	Perspicuity, Useful, Trustworthiness, Content Quality	Novelty	Clarity (interface) and having a result	Length	Akinator
6	Good	No	Bad	Trustworthiness	Novelty	Clarity (interface)	Stuck in a wrong path with no way out	Internet research
7	Good	No	Good	Trustworthiness	Attractiveness and Novelty	Clarity (interface) and easy to use	Lack of novelty	No
8	Sufficient	Yes	Good	Usefulness, Trust, Content Quality	Attractiveness and Novelty	Concept of/idea behind the product, objectivity, relevant	Language (English)	No

						questions and answers		
N	**English level**	**Covid check**	**Path**	**Most important**	**Least important**	**Liked**	**Disliked**	**Similar app**
9	Sufficient	No	Bad	Content Quality	Not mentioned	Clarity (interface) for questions	Clarity (interface) for result page, and Language (English)	No
10	Good	No	Good	Trustworthiness, Content Quality	Attractiveness and Novelty	Clarity (content)	Slow interface	No
11	Good	No	Good	Trustworthiness	Attractiveness and Novelty	Result page: It tells you to go to a doctor and ability to better understand the process	Length	Yes
12	Sufficient	No	Bad	Perspicuity and trustworthiness	Attractiveness and Novelty	Exhaustive result page and many functionalities	Attractiveness	Akinator
13	Good	No	Good	Trustworthiness	Not mentioned	Thoroughness	Limited symptom input	No
14	Sufficient	No	Bad	Trustworthiness, Content Quality	Attractiveness	Available choices	Explanations difficult to understand	Yes
15	Sufficient	Yes	Good	Trustworthiness	Attractiveness	Close-ended questions and efficiency	Question order	No
16	Sufficient	No	Bad	Trustworthiness, Content Quality, Efficiency	Attractiveness and Novelty	Quality	Language (English)	No
17	Sufficient	No	Bad	Trustworthiness and Efficiency	Not mentioned	Easy to use	Limited symptom input, close-ended, incomplete	Doctissimo

N	English level	Covid check	Path	Most important	Least important	Liked	Disliked	Similar app
18	Good	No	Bad	Usefulness, Dependability, Content Quality	Attractiveness and Novelty	Concept of/idea behind the product	Result page: not having a result at the end	No
19	Sufficient	Yes	Bad	Efficiency and Content Quality	Not mentioned	Contextual utility and relevance	Nothing	No
20	Good	No	Good	Trustworthiness and Efficiency	Attractiveness	(Logical) order of question	Missing “I don't know” options	Yes
21	Sufficient	No	Good	Perspicuity and efficiency	Not mentioned	Result page	Similar questions	No
22	Good	No	Good	Trustworthiness	Attractiveness	Thoroughness	Limited symptom input	Internet research
23	Good	No	Good	Trustworthiness, Content Quality, Dependability	Not mentioned	Easy to use	None	No
24	Good	No	Good	Trustworthiness, Dependability	Attractiveness	Plausible results	None	Doctissimo
25	Good	No	Good	Usefulness, Content Quality, Trustworthiness	Attractiveness and Novelty	Thoroughness	Slow	Internet research
26	Good	No	Good	Perspicuity, Trustworthiness	Attractiveness and Novelty	Clarity of content and interface	Unclear words and symptoms missing	Health insurance app
27	Good	No	Good	Dependability, Trustworthiness	Novelty	Simplicity and Clarity (Interface)	Small result page	No

				and Perspicuity				
28	Good	Yes	Good	Trustworthiness and Perspicuity	Attractiveness	Concept of/idea behind the product	Similar questions	Insurance websites
29	Good	No	Good	Content Quality and Trustworthiness	Attractiveness	Easy to use	None	No

9.5.8. Interview transcriptions

Participant 1

Niveau d'anglais annoncé : B2

Maitrise de la langue : Suffisant

Trajet dans l'application : Mauvais

Commentaire : le participant venait de se réveiller

Interviewer	*Wait, wait* ! *Do you prefer English or French* ?
Participant 1	Comment ?
Interviewer	Tu préfères français ou anglais ?
Participant 1	Les deux.
Interviewer	Il n'y en a pas un dans lequel vous vous sentez plus confortable ?
Participant 1	Pour moi, le plus important c'est les premières cinq. Mais *it says* « *place at most eight answers ». So*, c'est pour cela que j'ai choisi les huit.
Interviewer	Et pourquoi est-ce que vous trouvez que ces qualités sont importantes ?
Participant 1	Pourquoi c'est important ? La confiance, en fait, si on utilise une application nous qui va m'aider avec ma santé, c'est très important pour moi de... [rire] d'avoir l'impression d'être en confiance avec l'application, c'est ma santé, quoi. Je ne sais pas si tu... Si vous voyez.
Interviewer	Hm-hm.
Participant 1	Et la confiance, pour moi, c'est le plus important. Avec le contenu. C'est des enjeux délicats, vous voyez ?
Interviewer	Ouais.
Participant 1	Alors avec le *content*, c'est dans l'ensemble, la qualité, la confiance demande la qualité du contenu. Alors, normalement, maintenant, si j'ai un problème de santé, je ne m'intéresse pas vraiment à l'interface, si elle est jolie ou pas, je pense que la chose la plus important c'est que, à la fin, je puisse savoir ce que j'ai, non ? C'est un docteur comme on disait, mais si l'application essaie... [rire] les contenus sont de qualité et c'est efficient pour moi, voilà.
Interviewer	Ouais.
Participant 1	Après, bien sûr, si est clair, c'est vraiment important parce que je pense, par exemple, que si ma mère, elle a une maladie et que j'utilise l'application et qu'elle l'utilise, j'aimerais bien qu'elle puisse comprendre l'information. Et bien sûr, que si on a une interface agréable c'est, c'est joli, non ? Mais [bruit qui désigne qu'il ne sait pas] voilà.
Interviewer	Et comment est-ce que vous décrirez à votre expérience avec ce produit ?

Participant 1 En fait, c'est bizarre parce que [rire] je n'ai pas pu [pi], je viens de me réveiller [rire] et c'est difficile de jouer le rôle de, voilà, d'une personne avec une allergie, mais ce que je trouve c'est que c'est long. C'est vraiment lent d'utiliser.

Interviewer OK.

Participant 1 Je ne sais pas si c'est [u]n effet [lié au fait que] je me suis réveillé [rire] il y a quelques instants, mais pour moi c'était trop long.

Interviewer D'accord.

Participant 1 Mais, c'est ce qui est bien trouvé, j'ai trouvé, que les petites icônes avec l'information quand je ne sais pas qu'est-ce que c'est la maladie, la caractéristique, c'est vraiment utile. Je vous ai demandé deux fois ce que veut dire quelque chose, vous voyez ?

Interviewer Hm-hm. Ouais.

Participant 1 Et c'est agréable, en fait. Non, j'ai beaucoup aimé mais, mais ! C'est vrai que c'est trop lent. Mais, ouais, c'est trop lent.

Interviewer Est-ce que le produit correspondant à vos attentes ?

Participant 1 [pause réflexive] Je n'avais pas une attente ? [rire]

Interviewer OK.

Participant 1 Ouais, non.

Interviewer OK. Et qu'est-ce que vous avez le plus apprécié dans ce produit ?

Participant 1 Pardon ? Le plus apprécié ? Les petites icônes avec l'information. Et je trouve aussi que limiter les choix en petit [pi] en liste, c'est cool, parce que normalement, quand on va chez le médicin, euh, médecin, désolé [rire], je suis hispanophone [rire] quand on va chez le médecin, il te demande « qu'est-ce que vous sentez ? » et tu ne sais rien dire. Ici...

Interviewer Hm-hm.

Participant 1 *You can feel a bit related* avec le, le système -

Interviewer OK.

Participant 1 - d'analyse. Et c'est chouette parce que ça t'aide à exprimer que je save pas... je ne sais pas [rire] comment exprimer [rire].

Interviewer Hm-hm. OK.

Participant 1 Voilà, c'est ça.

Interviewer Et qu'est-ce que vous avez le moins apprécié dans ce produit ?

Participant 1 C'est trop lent.

Interviewer Hm-hm. OK.

Participant 1 Et moi j'aime bien prendre du temps, mais c'est trop long.

Interviewer OK. Et est-ce que quelque chose a fonctionné de manière inattendue ?

Participant 1 Je ne pense pas, non.

Interviewer	Est-ce que vous aviez utilisé une application similaire auparavant ?
Participant 1	Par rapport à la santé, non. Pas du tout.
Interviewer	D'accord, OK. Bon, je vous laisse continuer alors. Vous pouvez arrêter le partage d'écran.

Participant 2

Niveau d'anglais annoncé : B1-B2

Maitrise de la langue : Suffisant (à peine)

Trajet dans l'application :

Commentaire : le participant a traduit certains items depuis une application de traduction sur son téléphone.

Transcription de l'entretien

Interviewer	Attendez. [rire] Pourquoi vous les aviez hiérarchisés de cette manière ?
Participant 2	Pour quoi je les ai hiérarchisés comme ça ? Parce que je me doutais [racle la gorge] comment dire que je me doutais, c'est pas du tout ça que je voulais dire. C'est clairement comme ça que je voyais les choses quand je les ai faites et en gros j'ai mis les adjectifs que je pensais en premier, en premier, et après je suis descendu dans ce que je pensais le plus en plus, qui s'éloignait de mon propre avis je les ai mis en dernier.
Interviewer	Ouais. Pourquoi est-ce que ces qualités sont importantes pour vous dans une application de santé ?
Participant 2	Ah bah clairement parce que ça touche à la santé et du coup on ne peut pas mettre tout et n'importe quoi et on doit croire aux données qui émergent de là pour pas que le site nous dise n'importe quoi. Par exemple là, j'avais mal à l'oreille, c'était l'exemple. Bah, ils ne vont pas me dire que j'ai fait une gastro. C'était logique dans la logique des choses. Pour moi, tout ce qui sort de là, puisque ça touche à la santé, doit être logique et au moins sembler vrai, quoi.
Interviewer	OK. Comment décririez-vous votre expérience de cette application ?
Participant 2	Au départ, je me demandais ce que je devais faire, et en gros je n'ai pas trouvé comment mettre plusieurs symptômes. Je ne sais pas si vous avez vu, mais j'ai un peu cherché, mais je n'ai pas trouvé comment mettre plusieurs symptômes, du coup, mais ça c'était un peu plus ennuyant. Mais sinon, c'était facile à comprendre, une fois que j'ai mis un symptôme, il suffisait de cliquer par rapport à ce que j'avais, et ça c'était super facile. Facile et rapide à faire, on va dire, si on est vraiment malade et on cherche un mini diagnostic c'est facile à faire.
Interviewer	Est-ce que le produit correspondait à vos attentes ?
Participant 2	Oui, franchement, c'était sobre, j'avais peur que ça soit quelque chose avec, comment dire, j'ai déjà vu des sites de santé avec sur le côté plein d'onglets et de choses comme ça. Et là c'était simple, il suffisait juste de

répondre, et il y avait que ça à l'image, et du coup on était bien concentrés dessus.

Interviewer — Qu'est-ce que vous avez le plus apprécié dans ce produit ?

Participant 2 — La simplicité.

Interviewer — Et le moins apprécié ?

Participant 2 — Que c'était simple et efficace.

Interviewer — Ouais, hm-hm.

Participant 2 — Moins apprécié ? C'est... J'ai toujours un doute, quand je suis sur des sites, de savoir d'où viennent les data et donc je ne sais pas trop si les diagnostics à la fin sont vrais, et c'est ça que j'apprécie un peu moins. C'est que j'aurais un diagnostic, mais je ne sais pas vraiment si c'est le vrai diagnostic que j'aurais d'un médecin, quoi.

Interviewer — OK. Est-ce que quelque chose a fonctionné de manière inattendue ?

Participant 2 — Non, à part comme j'ai dit, je n'ai pas trop trouvé comment mettre plusieurs symptômes.

Interviewer — Et est-ce que vous avez utilisé une application similaire auparavant ?

Participant 2 — Ce n'est pas vraiment similaire, mais je suis déjà allé voir, par exemple, si j'ai un mal de tête et un mal de bras en même temps, je vais taper ça sur internet et je vais tomber sur des sites comme Doctissimo ou des choses comme ça.

Interviewer — OK.

Participant 2 — Mais c'est plutôt moi qui tape mes symptômes et de là j'ai directement un diagnostic. Alors qu'ici, ça va quand même un peu plus loin, mais ça parait un peu plus... Un peu mieux quoi.

Interviewer — Super. Merci beaucoup. Vous pouvez arrêter le partager l'écran et terminer tranquillement le *survey*.

Participant 3

Niveau d'anglais annoncé : C1

Maitrise de la langue : Bonne

Trajet dans l'application : covid-check

Commentaire : Le participant n'avait pas accès à une webcam la première fois, donc nous avons reprogrammé.

Transcription de l'entretien

Interviewer	Pourquoi est-ce que tu les as hiérarchisés de cette manière ?
Participant 3	Tu veux que je retourne en arrière ?
Interviewer	Non, non.
Participant 3	Parce que [hésite, pause].
Interviewer	Pourquoi est-ce que ces qualités sont importantes pour toi ?
Participant 3	Pourquoi ?
Interviewer	Ces qualités sont importantes pour toi dans une application santé.
Participant 3	Parce que... Ah, OK, d'accord. Je sais que j'ai mis en premier, je pense que si je me rappelle bien, la confiance, *trustworthy*, quelque chose comme ça.
Interviewer	Oui.
Participant 3	Pour moi c'est la base quand on parle d'une application santé, la confidentialité et la confiance c'est le plus important, du coup j'ai commencé par mettre ça. Mais je peux mettre en arrière pour que je me rappelle du *ranking* que j'ai mis ? Parce que...
Interviewer	Je ne pense pas que tu puisses retourner en arrière. Je vais voir si je peux changer.
Participant 3	Je sais que j'ai mis vers la fin « *Novelty* ».
Interviewer	Oui.
Participant 3	Parce que je sais que depuis le covid, à la base, je ne suis pas très application santé, tout ça, je ne m'y connais pas trop.
Interviewer	Hm-hm.
Participant 3	Je sais qu'ils ont développé depuis. Je crois bien, ouais, du coup j'ai mis plus vers le bas. Je sais que vers le haut aussi j'ai mis ce qui est par rapport à l'efficience, tout ça, parce que j'estime que c'était efficient. Et je me rappelle plus ce que j'ai mis.
Interviewer	Facilité à-
Participant 3	Ah, oui, aussi, j'ai mis aussi « *Dependency* » vers la fin parce que c'est pas du tout dépendant, enfin, je pense que c'est pas du tout dépendant des individus, des personnes.
Interviewer	Hm-hm.
Participant 3	Des personnes qui décident de participer à ça. C'est une dépendance mutuelle, c'est comme si c'était, enfin, je ne

sais pas comment expliquer, ça, mais pour moi c'est vers la fin, quoi.

Interviewer Hm-hm.

Participant 3 Enfin, les personnes ont besoin de l'application et l'application a besoin des personnes. Du coup, ce n'est pas forcément une dépendance à sens unique, c'est les deux. Du coup, les deux... Si les deux ont le choix de pas participer ou de ne pas appliquer, chacun son problème. Mais c'est une dépendance mutuelle, en fait. Du coup la dépendance mutuelle, c'est l'indépendance. La dépendance individualisée, quoi. Et du coup, il y a quoi encore que j'avais mis. J'avais mis... J'ai oublié.

Interviewer OK. Mais en tout cas tu m'as donné plein d'informations, je comprends mieux pourquoi tu as catégorisé les choses que tu les as faites.

Participant 3 Hm-hm. Super. Merci.

Interviewer Et comment ce que tu décrirais ton expérience de l'application ?

Participant 3 De l'application ? C'était super, en fait, c'était super, je trouvais que c'était super clair. Je sais que moi, personnellement, moi, je n'aime pas quand il y a trop de chichis, beaucoup charabia, beaucoup de d'explication pour dire que « peut-être que c'est ça, peut-être que ce n'est pas ça ». Je préfère directement *go straight to the point* et tu me dis les symptômes, que ça correspond à quoi et ce que je dois faire. Du coup, j'ai *enjoy* l'expérience sur l'application et voilà.

Interviewer Est-ce que le produit correspondant à tes attentes ?

Participant 3 Ouais. Ouais.

Interviewer Est-ce que tu avais des attentes ?

Participant 3 Je n'ai pas forcément pensé aux attentes, mais quand tu me pose la question, je me dis que, oui, au final, oui. Quand je réfléchis.

Interviewer Qu'est-ce que tu as le moins apprécié dans ce produit ?

Participant 3 Dans le projet ?

Interviewer Dans le produit.

Participant 3 Dans le produit. [pif]

Interviewer Ouais.

Participant 3 Je n'ai rien pas apprécié. J'ai apprécié, en fait.

Interviewer Ouais.

Participant 3 Je dois absolument chercher une critique ?

Interviewer Non.

Participant 3 Ouais, OK.

Interviewer Non, si t'en as pas, tu n'en as pas.

Participant 3 [rire] Je n'ai pas de truc négatif.

Interviewer Est-ce que tu avais déjà utilisé une application similaire auparavant ?

Participant 3 Jamais.

Interviewer OK.

Participant 3 Jamais.

Interviewer	Et est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 3	Comme je n'avais pas d'attentes, je suivais juste le processus, du coup, je n'ai pas forcément pensé à ce qui était attendu et inattendu. Peut-être, mais c'est par rapport au Limesurvey, j'avais râle... mais ce n'est pas la question que tu poses ?
Interviewer	Non.
Participant 3	Non. Je n'y avais pas forcément pensé, j'allais juste au compte-gouttes et je faisais des choses, quoi.
Interviewer	OK. Top. Merci beaucoup. Tu peux arrêter le partage et terminer tranquillement dans le *survey*. Ce n'est pas fini, juste le partage, oui.

Participant 4

Niveau d'anglais annoncé : B2

Maitrise de la langue : Bonne

Trajet dans l'application :

Commentaire : le participant n'est pas venu la première fois. Il a reprogrammé après que je lui aie envoyé un message le jour même.

Transcription de l'entretien

Interviewer	Pourquoi est-ce que vous avez hiérarchisé -
Participant 4	Bah il faut -
Interviewer	Pourquoi est-ce que -
Participant 4	Ce n'est pas ce qu'il faut faire ?
Interviewer	Oui
Participant 4	Pardon ?
Interviewer	Non, non, c'est ce qu'il faut faire.
Participant 4	Ah, OK.
Interviewer	Je n'ai pas terminé, en fait, ma phrase [rire]. Pourquoi est-ce que vous avez hiérarchisés de cette manière ?
Participant 4	Uh.
Interviewer	Pourquoi est-ce que ces qualités sont importantes pour vous ?
Participant 4	En fait, en réalité, ce n'est pas facile pour moi, à part les deux dernières (Attractiveness et Novelty), ce n'est pas facile de les mettre d'une certaine manière. Je prends, les dernières, enfin, personnellement, je m'en... Enfin, ce n'est pas ce qui m'intéresse trop dans une application de santé.
Interviewer	Hm-hm.
Participant 4	Après, j'ai, allez, on va dire que celle-là (Usefulness), j'aurais pu la mettre tout à la fin, mais après ça serait un peu hypocrite, parce que si elle n'est pas utile, je ne vais juste pas la télécharger. Elle pourrait même être en premier, parce que les applications qu'on va télécharger, on va les télécharger pour une question d'utilité. Mais sinon en dehors de ça, c'est vrai que pour le reste [pause] Après peut-être que la perspicacité (Perspicuity) je mettrais un peu plus bas, si en y réfléchissant, parce que, à un moment, je trouvais que les questions partaient un peu en dehors de ce que j'avais. Enfin, par rapport au test. Mais après, quand après quand il y a eu les réponses, je me suis rendu compte de pourquoi il y avait ces questions-là. Mais sur le moment, c'est vrai que j'étais là « les questions sont un peu bizarres », mais c'est un peu plus dans cette idée-là. Mais j'ai un peu pris cette idée-là dans l'idée de perspicacité de l'application, à moins de la suivre sur le raisonnement, ça partait sur le côté. C'est pour ça que j'avais mis plus sur le haut dans un premier temps, mais

c’est que c’est plus par rapport au test. Mais je ne sais pas si ça va un peu avec efficience. Perspicacité c'est un peu cette idée-là. Mais c'est vrai que mettre avec efficience serait plus logique.

Interviewer OK. Comment décririez-vous votre expérience de Babylon Health ?

Participant 4 C'est un truc que j'ai fait, enfin, moi personnellement, aller sur des applications et tout ça pour me médicaliser, personnellement, ce n’est pas un truc que je ferais, premier point. Donc c'est un peu une découverte de cette pratique. Après c’est vrai que dans des cas comme ici avec la pandémie, ça peut être intéressant, pour éviter un peu le contact ou pour... Après, ça m'a peut-être donné envie de l'utiliser, mais dans d'autres contextes, mais peut-être, parce que je suis chef scout, ça pourrait peut-être me servir pour voir si on doit aller au médecin ou pas, rapidement ou pas, en fonction des choses, enfin, que les rencontres pour nous pas de... Parce que, des fois on va chez le médecin, et le médecin nous dit, enfin il n’y a pas besoin de, enfin, ce n’était pas ultra nécessaire et tout, c'est vrai que là-dessus, c'est le point qui pourrait m’être le plus intéressant. Enfin, voilà, pour l'utilisation, c'est vraiment ce qui va en sortir d'intéressant. Mais sinon, je pense que dans ma vie de tous les jours, je ne pense pas que j'utiliserais régulièrement.

Interviewer Est-ce que le produit correspondait à vos attentes ?

Participant 4 [pause]

Interviewer Est-ce que vous aviez des attentes ?

Participant 4 Je n’avais pas vraiment d'attentes, conscientes en tout cas. Mais comme j'ai dit, à un moment, comme j'ai dit, j'avais l'impression que l'application partait un peu n'importe où. Mais, au final, quand on arrive à la fin, on comprend mieux pourquoi il y a eu ces questions. Enfin, pour mieux préciser. Au final, je pense que j'avais quand même une certaine attente qu'il devine plus vite peut-être ce que j'ai, ou quelque chose dans ce genre-là. En tout cas pour une première fois. Mais ce n’est pas quelque chose que j'avais au départ, c'est quelque chose que je me suis rendu compte en mettant... Enfin, voilà. Mais sinon je n’avais pas vraiment d'attente de base.

Interviewer OK. Qu'est-ce que vous avez le plus apprécié dans ce produit ?

Participant 4 Pardon, je n'ai pas bien compris la question.

Interviewer Qu'est-ce que vous avez le plus apprécié dans ce produit ?

Participant 4 [pause] La fin. La fin, je pense qu'à la fin j'ai apprécié le fait des différentes propositions qu'il faisait et la

	petite explication qui allait avec ce que ça recouvrait comme maladie et, d'une certaine manière, ça permettait de voir un peu pourquoi il avait posé telle ou telle question, pour essayer de cerner pourquoi on est malade et tout. Enfin, je trouvais ça bien la zone des résultats, la manière dont elle était mise en place.
Interviewer	OK. Et ce que vous avez le moins apprécié ?
Participant 4	Enfin, je vais revenir encore sur ce que j'ai dit, mais c'est ce qui m'a paru le plus bizarre, en tout cas, cette question qu'à un moment, voilà, j'avais l'impression que l'application était un peu perdue. C'est un peu débile, mais c'est un peu, je ne sais pas si vous voyez l'application Akinator, où on pose des questions. En fait, c'est une application où on doit, c'est une application de jeu, en gros, le but c'est d'essayer de deviner, enfin, une IA, essaie de deviner le personnage que vous avez en tête.
Interviewer	Hm.
Participant 4	Et il y a toujours un moment dans ces applications-là où le personnage, l'intelligence artificielle va, comme elle doit tester toutes les possibilités, elle va partir un peu sur des choses qui ne semblent pas super liées dans un premier temps, mais à la fin quand on arrive au résultat, on se rend compte qu'elle a posé ces questions pour éliminer toute une série de symptômes. Enfin, c'est un peu cette impression-là que j'avais. Enfin, je ne sais pas trop où ça allait. Enfin, je ne sais pas si je suis clair et si vous voyez ce que je veux dire.
Interviewer	Oui, oui, je comprends tout à fait.
Participant 4	Voilà.
Interviewer	Et est-ce que vous aviez utilisé une application similaire à celle-ci ?
Participant 4	Non.
Interviewer	D'accord. Voilà. Je vous laisse continuer et terminer le questionnaire. Est-ce que.... Vous pouvez maintenant arrêter le partage d'écran
Participant 4	OK.
Interviewer	Comme ça vous pouvez répondre tranquillement.

Participant 5

Niveau d'anglais annoncé : B2

Maitrise de la langue : Bonne

Trajet dans l'application : Bon

Commentaire : /

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ? Pourquoi est-ce que ces qualités sont importantes ? Pourquoi est-ce que vous avez les avez hiérarchisés de cette manière ?
Participant 5	Bah, le plus important, ce serait d'être clair. Parce que quand on recherche... des informations médicales ou des conseils [pause] J'imagine que souvent, dans un genre d'état de stress, on n'a pas peut-être pas envie de comprendre des choses trop compliquées. Le fait que le processus soit très clair et en fait, il l'est, c'est genre « où est-ce que tu as mal ? ». C'est des trucs très basiques. C'est le plus important, je pense. Et ensuite, je pense à « *Useful* » parce que [pause] Bon, là, on ne parle que du diagnostic, mais j'imagine que si on a « *Useful* » dans le sens ça te donne des trucs concrets, des étapes concrètes à suivre. C'est très, enfin, « *grounded* » tu vois ? C'est concret. Ensuite j'imagine qu'il y a la « *Content Quality* » qui est que... Je confonds un peu avec « *Trustworthiness of Content* », parce que voilà, j'imagine que d'avoir du contenu dans lequel on peut avoir confiance, c'est aussi assez important. Le « *Attractiveness* », bon, j'imagine que c'est moins important. Si on peut reprendre, si on peut reprendre la terminologie du discours d'UX, évidemment, le caractère hédonique ici est très, très important, même si j'imagine que les gens ont envie d'avoir une interface qui les rassure, j'imagine. Quelque chose qui est très épuré comme ça, juste clique là où tu as mal s'ils s'imaginent que ça a un côté assez efficace aussi. Donc, « *Attractiveness* », « *Efficiency* » ... « *Dependability* », qu'est-ce que c'était déjà ? « *Trustworthy and reliable* » [pf]... ça justement, c'est « *Content Quality* ». Je l'ai mis à la fin justement peut-être parce que je ne pensais pas le mettre.
Interviewer	D'accord. Est-ce que vous voudriez le déplacer ou ça vous convient comme ça ?
Participant 5	Ouf, ouais, vas-y on le déplace. Il est là, quoi ? Ouais, je ne sais pas vraiment faire la différence entre « *Content Quality* », « *Dependability* » et « *Trustworthiness of Content* ». Enfin, dans l'entier parce que...

Interviewer Ce sont des concepts très liés, donc c'est normal.

Participant 5 Hm-hm, ouais. Et enfin « *Novelty* » parce que j'imagine utiliser la clé de détermination. Ce n'est pas très, très neuf. Ça fait un peu penser à Akinator, donc ouais, juste cliquer sur des questions qui se suivent. J'ai répondu un peu plus tôt. Parfois, c'est un peu, ce n'est pas très, très clair parfois, les sauts qu'il fait. Parce qu'au début, on parle de notre nez, puis, il s'est très intéressé à la gorge et à la bouche. Bon, j'imagine que c'est lié parce que ces syndromes ORL sont liés, mais enfin voilà. Bon... on se concentre sur le nez, puis autre chose, puis une autre chose. J'imagine que pour faire un diagnostic, c'est important de savoir tous les endroits où ça fait mal. Voilà, ce n'est pas pire.

Interviewer Et comment est-ce que vous décririez votre expérience de Babylon ?

Participant 5 C'est ça ici ? Ou c'est une question qui est liée au truc d'avant ?

Interviewer C'est un petit moment « interview ». Entre les deux.

Participant 5 OK. J'ai trouvé ça plus long que je m'y attendais. J'imagine que c'est une bonne chose d'être exhaustif [rire]. Ouais, un peu un peu déroutant sur la longueur et sur l'exhaustivité du truc et... À la fin, je n'ai pas fait très attention. Je n'ai pas fait très attention à ce qu'il nous a dit à la fin, mais je pense qu'il n'y avait que deux résultats.

Interviewer Hm-hm.

Participant 5 Voilà, je ne sais pas ce qu'il y a derrière, mais je me dis a de grandes chances que ce soit l'un des deux avec toutes les questions qui ont fait. Donc quelque part on a un genre de confiance. Parce que, j'imagine, je pense, je pense aux autres, aux autres genres de diagnostics sur lesquels on fait des blagues du genre ouais, ouais, « Ben, dis-tu le nez bouché ? C'est que tu as le cancer, quoi ». Donc là, j'imagine qu'on peut faire plus confiance à ça.

Interviewer Est-ce que le produit correspondait à vos attentes ?

Participant 5 Oui, oui. Ça correspondait aux attentes.

Interviewer Et qu'est-ce que vous avez le plus apprécié dans ce produit ?

Participant 5 [rire] La clarté, j'imagine, le fait que ce soit méthodique. Et le fait que ça donne un résultat à la fin. Bon, après je ne sais pas quelle est la véritable qualité de ce résultat à la fin. Ça ne donne que deux résultats à la fin, c'est là ça.

Interviewer Et qu'est-ce que vous avez le moins apprécié ?

Participant 5 Que ça soit long. J'imagine que ça peut être expliqué, mais, ouais, c'est assez long.

Interviewer Ok. Est-ce que quelque chose à fonctionner de manière inattendue ?

Participant 5	Non.
Interviewer	Et ce que vous aviez déjà utilisé une application similaire auparavant ?
Participant 5	Non, pas avec une clé de détermination comme ça. Avec des questions qui se suivent.
Interviewer	Ok. Je vais vous laisser alors terminer, le *survey* et vous pouvez arrêter le partage. Vous pouvez terminer tranquillement.

Participant 6

Niveau d'anglais annoncé : C1/C2

Maitrise de la langue : excellent

Trajet dans l'application :

Commentaire : le participant a rencontré quelques problèmes de connexion.

Transcription de l'entretien

Interviewer	Attendez.
Participant 6	Oui.
Interviewer	Ah, OK. Je suis désolée, l'écran passe beaucoup plus vite.
Participant 6	Oui.
Interviewer	Je voulais vous demander -
Participant 6	Ah, j'ai passé sans faire exprès [pause]
Interviewer	Pourquoi est-ce que ces qualités sont important [pause]
Participant 6	D'accord. A la précédente ?
Interviewer	Oui. Si vous voulez avoir les réponses sous les yeux. Je voulais vous demander pourquoi ces qualités sont importantes pour vous. Pourquoi est-ce que vous les avez hiérarchisés de cette manière ?
Participant 6	Pour moi, par exemple, quand c'est au niveau santé, même si ce n'est quelque chose qui est, enfin, si c'est l'application est super jolie, me donne super envie de l'utiliser, c'est un peu secondaire parce que si à la fin elle me donne des résultats... Enfin, de toute façon je devrais appeler un médecin, parce que c'est pas du tout sûr que les résultats soient bons. Enfin, s'ils me disent [Wi-Fi coupe].
Interviewer	Ah. Internet a encore coupé je pense.
Participant 6	Est-ce que ça remarche ?
Interviewer	Ca remarche
Participant 6	Je suis vraiment désolé, c'est mon internet. Aujourd'hui en plus... D'habitude ça va. OK. Je ne sais pas ce que tu as entendu ou pas.
Interviewer	Alors, les applications santé, le plus important c'est le contenu.
Participant 6	Voilà, c'est la confiance dans ce que, enfin, dans le fait que, il y ait quand même un, qu'on ait confiance dans l'application, quoi. Que si ça vient, d'un truc, d'une association de médecins, j'aurai plus confiance là-dedans que si c'est un n'importe qui qui l'a fait, par exemple. Ou si c'est recommandé par mon médecin traitant, j'aurai plus confiance à l'utiliser. Du coup, faut aussi que je trouve que si c'est rapide à faire, genre si ça me prend vingt minutes pour faire mon, allez, diagnostic, je suis peut-être mieux d'appeler mon

médecin traitant. Donc, il faut que ça soit quand même rapide, sans donner des résultats à l'arrache comme ça.

Interviewer Hm-hm.

Participant 6 Oui, le *Usefulness* je l'ai mis au milieu parce que j'ai envie que ça me donne des résultats, des choses qui me sont utiles. Genre « est-ce qu'il faut que tu appelles un médecin, ou que t'ailles aux urgences », ou quoi. Mais je préfère d'abord avoir confiance en tout ça, par exemple. Du coup, pour moi le côté « *Novelty* » et tout ça, ce n'est pas ce qui m'attire là-dedans. Enfin, ce n'est pas ça qui va me faire utiliser. Enfin, s'il y a deux applications qui ont la même, toutes les autres qualités les mêmes, alors oui, je vais aller vers celui qui a le petit plus côté attrait, niveau, genre, des choses en plus. Sinon, j'irais plutôt vers celle qui est la plus, enfin, laquelle je peux avoir plus confiance, quoi. Ouais. Enfin, voilà à peu près. Je ne sais pas s'il y a autre chose que...

Interviewer Oui, j'ai d'autres questions.

Participant 6 Ouais.

Interviewer Comment est-ce que vous décririez votre expérience de Babylon Health ?

Participant 6 Bah, j'avais un peu, fondamentalement... En fait, quand ça va, ça va. Mais, je sais que, moi, je suis quelqu'un qui n'est un peu jamais sûr de ce qu'on pose comme question, et du coup je vais toujours revérifier, et puis finalement, souvent, peut-être pas répondre « le bon truc », ou peut-être me dire « ah tiens, quel symptôme est le plus, le plus fort pour moi ». Voilà, ça j'ai un peu mal lu les consignes, c'est ma faute, mais dans la vraie vie, genre, je serai peut-être en mode « tiens, est-ce que c'est mon nez qui coule, est-ce que c'est la douleur oreilles », ou quoi. Et du coup, j'ai un peu l'impression qu'ici, que si je suis parti avec le mauvais truc au début, je pars dans des chemins qui m'éloignent en fait du « bon résultat », je n'ai pas l'impression que ça me corresponde. Et du coup ce qui est peu dommage, ouais, c'est que, du coup, j'ai complété pas mal de trucs, puis, en fait, c'est un peu perdu. Je dois recommencer parce que peut-être que, ouais, au début j'ai pris le mauvais chemin. Et donc, tout ce que j'ai noté n'a servi rien à la fin. Donc ça, je trouve ça un peu dommage donc c'est un peu le côté négatif du truc. Sinon, je pense que c'est quand même, enfin, si ça marche bien, ça peut être vraiment pratique. Mais bon, s'il me faut une demi-heure pour noter mes trucs, alors je préfère aller voir un médecin parce qu'il saura me dire « bah non, ton symptôme, là, ce n'est pas le truc le plus important ». Ou peut-être juste un bête truc ça serait, enfin, parce qu'il disait « tiens, s'il y a plusieurs symptômes, le plus

important », mais peut-être alors noter plusieurs symptômes à la fois dès début, pour avoir... Pour voilà, les gens un peu comme moi, pour qu'ils ne soient pas perdus, à devoir faire un choix dès le début, quoi.

Interviewer Ouais.

Participant 6 Enfin, voilà c'est un peu ce que ce que je pense, mais, sinon je pense que ça a un intérêt dans le futur. Genre, moi, je n'aime pas sonner au médecin traitant mais, du coup, si à un moment j'ai un truc qui me permet de dire « bah, oui ça vaut la peine » ou pas que tu sonnes à ton médecin traitant ou « non, tu vas l'embêter, t'as probablement rien de vraiment important ». Ça peut être super utile je l'utiliserai très certainement mais, ouais. Enfin, voilà.

Interviewer Est-ce qu'est le produit correspondait à vos attentes ?

Participant 6 Peut-être que dans... En fait, le résultat final, c'est plus ou moins ce à quoi je m'attendais. Mais, c'est peut-être le processus qui était un peu long et, ouais, un peu, du coup, un peu plus long que ce à quoi je m'attendais et un peu plus genre « moins tolérant à l'erreur » que ce que j'aurais espéré.

Interviewer OK.

Participant 6 En gros.

Interviewer Qu'est-ce que vous avez le plus apprécié dans le produit ?

Participant 6 La simplicité au niveau de, chaque fois une étape, on ne doit pas se dire « tiens, comment je dois compléter ça ? », c'est juste cliquer et puis c'est bon. Ça, c'est bien. Mais en plus, les résultats à la fin, je trouve ça pas mal. Surtout le fait de dire « bah oui, ça vaut la peine d'appeler un médecin » genre on ne dit quand même pas « OK, le jugement est... », enfin, « appelle quand même un médecin, au cas où, parce que ça peut être quelque chose d'important ». Enfin, je trouve ça bien que cela soit mentionné.

Interviewer OK. Et qu'est-ce que vous avez moins apprécié dans le produit ?

Participant 6 Le côté, du coup, j'ai l'impression d'être dans un arc de décision et de et que je peux vite partir à l'autre bout sans m'en rendre compte, quoi. Et ça je trouve un peu dommage.

Interviewer OK. Et est-ce que quelque chose a fonctionné de manière inattendue ?

Participant 6 Le choix au début, du coup, je ne m'attendais pas à ce que ça me lance dans ce, du coup dans un mauvais truc. Je me disais, tiens, j'ai mis un certain symptôme, et puis on reviendrait sur le reste. Et puis, en fait, ça a peut-être un peu trop rétréci la possibilité de ce que j'avais comme, du coup, trucs et finalement, je suis passé à

	côté de ce que j'avais vraiment comme problème, quoi, en gros. Je ne sais pas si je suis très clair.
Interviewer	Si, si.
Participant 6	OK, OK.
Interviewer	Et, est ce que vous aviez déjà utilisé une application similaire auparavant ?
Participant 6	Je ne pense pas. Peut-être le plus proche c'est taper sur internet ce qu'on a et puis avoir trente résultats différents avec un truc, c'est on a un cancer, l'autre c'est on a juste un rhume. C'est peut-être le plus proche que j'ai fait. Aussi, dans notre truc, le plus proche c'est juste « réserver leur médecin » mais par internet, quoi. Mais en vrai, en liant ça au fait que, je sais que mon médecin traitant, du coup, je prends rendez-vous par internet mais, avoir avec cette application-là, direct, oui. Enfin, je n'ai peut-être pas manqué ça dans la page de fin, genre la possibilité de direct prendre rendez-vous suite à ça avec son médecin traitant où on lui donne tout ça, pourrait être pas mal.
Interviewer	OK.
Participant 6	Genre. Enfin, voilà.
Interviewer	Super. Merci beaucoup. Je vais vous laisser terminer le questionnaire tranquillement.

Participant 7

Niveau d'anglais annoncé : B2

Maitrise de la langue : Bonne

Trajet dans l'application : Excellent

Commentaire : pas de commentaire

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ? J'aimerais vous poser quelques questions. Ah, il y a un bouton « *back* » en dessous. Enfin, « *previous* ». Pourquoi est-ce que vous les avez hiérarchisés de cette manière ? Pourquoi est-ce que ces qualités sont importantes pour vous ?
Participant 7	Je pense que sur un site au niveau de la santé, c'est plus important le contenu que la façon dont il est mis en page. Du coup, c'est pour ça que j'ai mis tout ce qui est nouveauté, l'attractivité du site, etc. en dernier, et plus « est-ce qu'on peut faire confiance au contenu », la qualité en premier.
Interviewer	Hm-hm. Comment est-ce que vous décririez votre expérience de Babylon Health ?
Participant 7	C'est très facile d'utilisation. C'est intuitif. On voit aussi quand les questions vont finir, car il y a une petite barre au-dessus, c'est très bien. Sinon, au niveau créativité du site, ce n'est pas... Il n'y a rien de nouveau. C'est utile on va dire comme site, mais après, faut toujours se méfier de mettre ses symptômes sur internet. On ne sait pas si on va nous donner des choses correctes ou pas. Donc c'est bien qu'eux, ils écrivent « consultez quand même un docteur ». On voit que c'est de qualité par rapport à quand on met nos symptômes sur Google. Sinon, facile d'utilisation surtout et intuitif.
Interviewer	OK. Est-ce que le produit correspondait à vos attentes ?
Participant 7	Ouais. On a des symptômes, on les rentre, et après on a des réponses. On nous dit « ah c'est peut-être ça ou c'est peut-être ça parce que vous avez ça ». Donc je trouve ça [pi] de l'utilisateur.
Interviewer	Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 7	Pardon ?
Interviewer	Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 7	La facilité d'utilisation.
Interviewer	Et ce que vous avez le moins apprécié ?
Participant 7	[pause] La mise en page du site, là. Ouais, la créativité, la nouveauté. Enfin, c'est très simple, quoi.
Interviewer	Est-ce que quelque chose a fonctionné de manière inattendue ?

Participant 7	Non.
Interviewer	Et est-ce que vous aviez utilisé une application similaire auparavant ?
Participant 7	Non.
Interviewer	OK. Je vous laisse continuer tranquillement. Vous pouvez arrêter le partage d'écran.

Participant 8

Niveau d'anglais annoncé : B1

Maitrise de la langue : Suffisant

Trajet dans l'application :

Commentaire : le participant est arrivé un peu en retard car ZOOM n'était pas téléchargé et par conséquent était stressé.

Transcription de l'entretien

Interviewer	Attendez.
Participant 8	Ah !
Interviewer	Est-ce que vous pouvez revenir en arrière ?
Participant 8	OK. Oui ?
Interviewer	Pourquoi est-ce que vous les avez hiérarchisés de cette manière ?
Participant 8	Parce que du coup je pense que l'application est vraiment super intéressante. Je préfère ça que d'aller voir sur un blog ou regarder via Google et tomber sur des sites comme, je sais plus comment ça s'appelait, dont on nous dit qu'on va mourir dans les trente secondes. Voilà. [rire]. Je trouve efficace, quand même, si ce n'est, voilà, l'anglais, ou pour moi, c'est un peu plus compliqué. Je pense que le faire version francophone, voilà, ou l'adapter en tout cas pour les différentes langues, ça serait bien parce que je pense que c'est un très bon outil. Attractif, je trouvais qu'il était joli, chouette à utiliser, c'était aussi assez simple. Ce n'est pas non plus quelque chose d'innovant, parce que ça reste un petit questionnaire où on coche. Mais il ne faut pas plus, et je ne pense pas qu'une autre méthode aurait été plus intéressante, étant donné que de toute façon, c'est des questions comme le médecin le ferait, j'ai l'impression. Donc [pi] bien, bien m'installer. Du coup, les informations sont de qualité, les informations de qualité, enfin le contenu est de qualité. Ben, je trouvais que oui. Les questions sont pertinentes par rapport aux réponses qu'on donne précédemment et ça, enfin, ce sont des questions vraiment que le médecin pourrait poser, donc pour moi ça reste de la qualité. Maintenant, pour ce qui est des résultats, voilà, c'est pour ça le « *Trust* », je l'ai mis à la fin parce que les résultats, c'est à vérifier avec un médecin. Pour moi, ça reste quand même une figure d'autorité et j'aurais davantage confiance. Mais je trouve que c'est un bon début pour vérifier si ça nécessite quelque chose directement, ou si on peut laisser patienter, ou voilà. Perspicacité c'était en gros la rapidité, c'est ça ? En gros si les réponses sont ?
Interviewer	Claires et compréhensibles.

Participant 8 Ouais, alors je l'aurais peut-être mis avant.

Interviewer Vous les avez hiérarchisés en fonction de Babylon et pas des Health Apps en général ?

Participant 8 Hm-hm.

Interviewer D'accord. Ce que je cherchais à savoir, c'était qu'est-ce que pour vous, dans les applications santé, qu'est-ce qui est le plus important. Comment vous les auriez hiérarchisées par rapport à ça,

Participant 8 OK.

Interviewer Et pas par rapport à Babylon, spécifiquement.

Participant 8 OK. D'accord. Je pensais que c'était vraiment par rapport à mon expérience avec l'application. D'accord. OK. Donc, effectivement, pour moi ça va être quand même, que ça soit utile, puis la confiance, le contenu de qualité, quand même, hop, et après l'attractivité c'est en dernier et le fait que ça soit innovant aussi. Donc pour moi, il faut quand même que ça soit plus intéressant et que ça soit des contenus de qualités, sérieux, auxquels on peut faire confiance.

Interviewer OK.

Participant 8 Voilà.

Interviewer Comment décririez-vous votre expérience de Babylon Health ?

Participant 8 Voilà, c'est ce que je disais. Donc, en gros, je trouve que c'est vraiment une chouette application. Je ne sais pas si elle existe sur téléphone. Mais elle est simple d'utilisation, si ce n'est, pour moi, voilà, l'anglais, qui est pour moi un point faible, mais bon. Ça reste quelque chose qui dépend des gens. Ceci dit, j'ai assez compris, mais ça reste des « termes de médecine », donc ce n'est pas toujours accessible pour tout le monde, moi la première. Mais sinon, je trouve ça vraiment chouette. Mais voilà, ce n'est pas innovant, mais je ne sais pas ce qu'on aurait pu faire mieux pour poser les questions. Enfin, innovant dans le sens comment ça s'organise. Parce que l'idée est super innovante, je trouve ça génial. Mais, comment ça s'organise, voilà, ça reste un question-réponse. Maintenant, je ne vois pas comment on pourrait faire mieux. C'est super bien et je trouve que c'est joli à regarder, ça a l'air assez ins- en tout cas, ça inspire la confiance. Et ça prévient à chaque fois que ce n'est pas une solutions miracle et qu'il faut appeler le médecin, et donc ils font assez de sensibilisation là-dessus. Donc, franchement, je trouvais ça chouette. J'étais déjà assez impressionnée, dès le début, quand j'ai vu l'utilité de l'application, en fait. Donc je pars déjà avec un point très positif, parce que je me suis dit, « ouah, OK, c'est vraiment chouette ». Je préfère ça à regarder sur Google, quoi.

Interviewer	Est-ce que le produit correspondait donc à vos attentes ?
Participant 8	Oui. Ouais, ouais. Clairement.
Interviewer	Qu'est-ce que vous avez le moins apprécié dans ce produit ?
Participant 8	Je vais le redire, hein, la version anglophone. Sinon, je ne saurais pas dire, parce que franchement c'était facile d'utilisation, il y avait une bonne prise en main et c'était, en tout cas sur l'ordinateur, c'était joli, bien présenté, donc, voilà. Le contenu, les questions sont assez précises, il y a assez de choix. Peut-être rajouter une version « autre », mais en tout cas un petit truc à cocher « autre » s'il y a d'autres choses encore, ou peut-être plus de questions, je ne sais pas. Parce que c'est vrai que ça a été quand même assez rapidement. Mais j'ai répondu aussi que je n'avais pas d'autres symptômes. Je ne sais pas si ça aurait changé si j'avais répondu autrement. Mais, non, je trouve vraiment chouette. J'étais impressionnée dès le début [rire].
Interviewer	Et qu'est-ce que vous avez le plus apprécié dans le produit ?
Participant 8	Du coup, c'est l'idée en elle-même. Le produit en lui-même, ça inspire bien plus confiance, et on n'a pas les avis d'une personne lambda qui va nous dire tout de suite de solutions ou des problèmes un peu à la mort-moi le nœud. Donc là, on est vraiment sur des questions objectives et des réponses objectives qui nous sont fournies en fonction de ce qu'on dit. Donc, ça m'inspire beaucoup plus confiance, donc je pense que ça serait ça que je donnerais. L'idée en elle-même et que ça inspire plus confiance.
Interviewer	Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 8	Si ce n'est dernière partie, enfin, ici le *survey*. Sinon, non, dans l'application tout s'est bien déroulé. Je n'ai pas été étonnée de quelque chose, quoi que ça soit.
Interviewer	OK. Et est-ce que vous aviez déjà utilisé une application similaire auparavant ?
Participant 8	Pas du tout. Alors, là. Non.
Interviewer	Ça va. Je vais vous laisser continuer et terminer tranquillement le questionnaire.
Participant 8	Merci.
Interviewer	Vous pouvez arrêter le partage.
Participant 8	Ca va [rire].

Participant 9

Niveau d'anglais annoncé : B1

Maitrise de la langue : suffisant (à peine)

Trajet dans l'application :

Commentaire : le participant avait beaucoup de questions de vocabulaire.

Transcription de l'entretien

Interviewer	Pourquoi est-ce que vous les avez hiérarchisés de cette manière ?
Participant 9	[expire] Parce que je sais que l'expérience utilisateur, tout ce qui est l'aspect, que ça soit pratique et tout, est super important, mais le contenu de l'application prime. Surtout quand ça parle de santé.
Interviewer	Hm-hm.
Participant 9	C'est moins important si on parlait de vêtement, par exemple. Là, clairement le produit de vêtement est moins « vital ». Donc, clairement, j'aurais mis tout ce qui est l'usabilité en premier, mais là vu que ça parle de santé, je trouve que c'est le contenu, ouais, la qualité du contenu qui est le plus important.
Interviewer	Hm-hm.
Participant 9	Voilà.
Interviewer	Comment décririez-vous votre expérience de Babylon Health ?
Participant 9	Bah, j'ai été assez déçue, surtout la fin, en fait. Les questions ne sont pas mauvaises, sont même bien, mais moi j'aurais mis, après c'est un truc automatique, mais j'aurais mis des « autres », par exemple, pour pouvoir décrire. Mais bon, vu que c'est une application, ce n'est pas un médecin, donc, il n'y a peut-être pas des gens derrière. Certainement pas, en fait. Et donc j'aurais mis des « autres ».
Interviewer	OK.
Participant 9	Pour pouvoir décrire. Et sinon, à la fin, ils ont juste parlé du coronavirus. Et moi, je n'avais pas remarqué que les petits machins en dessous, je croyais que c'était pour plus d'informations, aller là-dessus, pour plus d'informations, aller là-dessus. Donc je n'ai même pas lu, en fait. Et c'était là le résultat. Je pensais que ça allait être plus *obvious*, plus grand, en mode « potentiellement, il y a ça, il y a ça, il y a ça » et, non. Là, il n'y avait pas. C'est plutôt la fin qui m'a été décevante.
Interviewer	D'accord. Donc, est-ce que le produit correspondait à vos attentes ?
Participant 9	Au niveau de la démarche, oui, ça va. Mais au niveau des résultats, non.

Interviewer	OK. Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 9	J'ai bien aimé la facilité avec laquelle on peut cliquer. Parce que l'anglais n'est pas ma langue maternelle, donc j'ai eu plus de mal avec ça. Mais si ça avait été en français, ouais, la facilité à lire, à cliquer, à savoir très vite. On n'y va pas par quatre chemins, donc c'est bien.
Interviewer	OK.
Participant 9	Voilà.
Interviewer	Et ce que vous avez le moins apprécié ?
Participant 9	Le moins apprécié ? La dernière page, la page des résultats, qui était, selon moi, pas bonne. Je n'ai pas vu directement mon truc, enfin, je n'ai pas vu directement mon résultat. On m'a directement parlé du coronavirus et des différents symptômes, alors que, je n'ai pas demandé pour le coronavirus ou quoi que ce soit.
Interviewer	Hm-hm
Participant 9	Donc, ouais.
Interviewer	Voilà.
Participant 9	Ce n'est pas encore fini, non ? J'ai encore des trucs ?
Interviewer	Oui, oui, oui.
Participant 9	Ah oui, OK.
Interviewer	C'est une petite interview entre les deux.
Participant 9	Ah, oui, OK. [rire]
Interviewer	Je rebondis un peu sur
Participant 9	Non mais c'est -
Interviewer	votre échelle.
Participant 9	D'accord.
Interviewer	Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 9	[ch-] Est-ce que quelque chose a fonctionné de manière inattendue ? [-ch] Non ? Je ne pense pas, non.
Interviewer	Et est-ce que vous aviez utilisé une application similaire auparavant ?
Participant 9	Non.
Interviewer	D'accord. Je vous laisse terminer tranquillement le questionnaire. Vous pouvez arrêter le partage d'écran.
Participant 9	Ah, d'accord. Ah, voilà. OK.

Participant 10

Niveau d'anglais annoncé : C2

Maitrise de la langue : Excellent

Trajet dans l'application : excellent

Commentaire : no comment

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ?
Participant 10	Oui.
Interviewer	Est-ce que vous pourriez me dire pourquoi est-ce que vous avez hiérarchisés les questions de cette manière ? En quoi est-ce que - pourquoi est-ce qu'elles sont importantes pour vous ?
Participant 10	Quand on parle de santé, c'est ce qui me semble le plus important, c'est de pouvoir faire confiance à l'information qu'on a, surtout si c'est un diagnostic. Même si ça ne remplace pas une visite chez le médecin, c'est quand même ce qui me semble le plus important. Après, effectivement, tout ce qui est de l'ordre de la confiance et du contenu, pour moi, c'est vraiment les qualités essentielles et puis, après, faut que ça soit utile. Après, tout ce qui est de la nouveauté ou de l'attractivité du produit, ça vient vraiment... Enfin, pour moi, ce n'est pas ce qui compte le plus. Faut d'abord que je puisse faire confiance aux informations que je peux trouver, avant de me dire, « ah, puis en plus c'est joli » ou « oh, tiens, c'est bien fait, puis c'est nouveau ».
Interviewer	OK. Comment décririez-vous votre expérience de Babylon Health ?
Participant 10	En fait, j'ai trouvé ça assez - un peu long. Bon, après, là, du coup, j'ai vu le temps que ça m'a pris, ce n'était pas si long. Mais c'est vrai que j'ai trouvé que les enchainements des différents écrans étaient un peu lents, je trouvais, par rapport par à ce qu'on peut avoir d'habitude sur internet. Du coup, surtout, je ne sais pas, il n'y avait pas tellement de barre de progression, on ne sait pas trop quand ça, on va avoir les résultats, combien de temps ça va durer. Sinon, j'ai trouvé que les informations étaient assez claires. Après, effectivement, le contexte du covid ralentit un peu le processus quand on vient autre chose que pour un diagnostic symptômes covid ou pas. Sinon, je trouvais assez clair, effectivement, que l'écran soi, enfin, qu'il y ait que les questions avec les symptômes et qu'il n'y ait pas d'autre informations parasites. Ça, j'ai trouvé que c'était clair. Mais, c'est vrai qu'on ne voit pas trop jusqu'où on va aller dans les questions qu'on va nous poser.
Interviewer	Est-ce que le produit correspondait à vos attentes ?

Participant 10	J'avoue que, du coup, je ne me suis pas trop projetée dans ce que je j'attendais du produit. Mais, pour un produit santé, c'est très moderne en fait, Enfin, par rapport à d'autres sites qui existent. En France, on a Doctissimo. Par rapport à ce type de produits, encore, Doctissimo, ils ont évolué en termes de mise en page du site, maintenant c'est plus moderne qu'avant. Mais, non, je trouve que c'est moderne par rapport à ce qu'on d'un site un peu de « diagnostic », on s'attendrait à une interface un peu plus vieille.
Interviewer	Qu'est-ce que vous avez le moins apprécié dans ce produit ?
Participant 10	Peut être la partie, justement, avant le diagnostic, où on passe, je sais plus combien d'écrans, mais au moins un ou deux, avant d'aller justement sur la partie diagnostic. Enfin, le fait que ça ralenti le fait qu'on arrive au diagnostic.
Interviewer	Hm-hm. Qu'est-ce que vous avez le plus apprécié ?
Participant 10	La clarté des informations.
Interviewer	Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 10	Le résultat, en fait, du diagnostic.
Interviewer	D'accord.
Participant 10	Enfin, d'avoir les différentes informations énoncées aussi clairement et d'avoir, justement, le diagnostic qui est le plus plausible, arrivé en bout de parcours.
Interviewer	Est-ce que vous aviez utilisé une application similaire auparavant ?
Participant 10	Non.
Interviewer	D'accord. Je vais vous laisser terminer tranquillement le *survey*. Vous pouvez arrêter le partage d'écran.

Participant retiré

Niveau d'anglais annoncé : B2

Maitrise de la langue : insuffisant

Trajet dans l'application : /

Commentaire : Le participant n'a pas réussi à comprendre les instructions ou à faire l'expérience. Les résultats sont invalides.

Transcription de l'entretien

Pas de transcription, car expérience invalidée.

Participant 11

Niveau d'anglais annoncé : B2

Maitrise de la langue : excellent

Trajet dans l'application :

Commentaire : le participant était un peu stressé et connaissait déjà un peu les applications de santé.

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ?
Participant 11	Quoi ? Je reviens en arrière ?
Interviewer	Oui, s'il vous plait.
Participant 11	OK.
Interviewer	Pourquoi est-ce que vous les avez hiérarchisés de cette manière ? Pourquoi est-ce que ces qualités sont importantes p- ?
Participant 11	Ah, ouais, OK. Pour moi, c'est que ça réponde à la question et qu'on ne reçoive pas n'importe quelle réponse qui correspondrait pas du tout, et qui dirait de ne pas aller chez le médecin, ou que ce n'est pas grave, ou que ça soit complètement à côté. Pour moi, ça c'est le plus important. C'est ce qu'on attend aussi quand on va chez le médecin. Et puis, j'ai mis, lequel.... « Dependa- », on, « *Perspicuity* », mais justement pour les gens qui ne vont pas spécialement chez le médecin, peut-être pour des raisons, je ne sais pas, pour que tout le monde puisse comprendre, quoi, en gros. Et ouais, après, ça, « *Novelty* », j'ai mis en dernier parce que j'ai déjà vu des app comme ça, il y a des... Je ne dirais pas, des années, mais probablement, quand même. Et voilà., Et « *Attractiveness* », c'est joli mais elles se ressemblent toutes quand même. Bon, à part que c'est un peu plus moderne et un peu plus coloré, peut-être, et pas de grande différence.
Interviewer	OK.
Participant 11	Voilà.
Interviewer	Et -
Participant 11	Bon, je continue ?
Interviewer	Non, pas encore.
Participant 11	Ah OK. Ouais ?
Interviewer	Comment décririez-vous votre expérience de Babylon Health ?
Participant 11	C'était pas mal. C'était facile à utiliser. Et, bien parfois, il y avait des questions, là on avait.... C'était facile de dire « oui/non », mais pas toujours, en fait. Parfois les questions ne sont pas évidentes à répondre, mais bon, elles ne sont pas compliquées non plus. C'était [pf] je ne dirais pas que c'était long, mais ça, ça allait, quoi. Ce n'était pas non plus... Au final, c'est ce qu'il faut à mon

	avis, pour essayer un peu d'éliminer un peu les possibilités, même si du coup il y a quand même beaucoup de questions. Mais, ça m'avait l'air efficace et les résultats m'ont paru. Moi, j'ai souvent des trucs aussi aux oreilles, donc ça m'a paru... je ne sais pas comment dire, qu'on peut avoir confiance, quoi.
Interviewer	Est-ce que ça correspondait à vos attentes ?
Participant 11	Ouais, c'est ce à quoi je m'attendais.
Interviewer	Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 11	Ca m'a beaucoup plut qu'ils disent quand même d'aller chez le médecin, qu'il y ait des conseils et qu'il y ait moyen de, je l'ai pas fait, mais qu'il y ait moyen de voir les explications sur les questions.
Interviewer	OK.
Participant 11	Il y avait un peu plus que juste les questions « Est-ce que vous avez mal aux oreilles, oui/non », « est-ce que vos oreilles saignent [rire] ou/non » [pif]
Interviewer	Qu'est-ce que vous avez le moins apprécié ?
Participant 11	C'était un peu long.
Interviewer	Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 11	Non, pas vraiment.
Interviewer	Et, donc, est-ce que vous aviez déjà utilisé une application similaire auparavant ?
Participant 11	Ouais.
Interviewer	Je vais vous laisser terminer tranquillement. Vous pouvez arrêter le partage.
Participant 11	OK. Oh, merde. Voilà.

Participant 12

Niveau d'anglais annoncé : B2

Trajet dans l'application : pas idéal

Maitrise de la langue : Suffisant

Commentaire : le participant parlait beaucoup et rapidement. Il n'a pas lu correctement les instructions à plusieurs moments.

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ?
Participant 12	Ah, oui, pardon. J'ai oublié quelque chose ?
Interviewer	Non, non. Je voulais vous demander pourquoi est-ce que vous les avez hiérarchisés de cette manière ?
Participant 12	Alors, « *Attractiveness* », c'est vraiment… En fait, l'attraction, ce n'est pas que ça m'a repoussé, c'est que cette idée que, j'ai trouvé que, enfin, c'est assez classique, en fait. Donc, ce n'est pas un défaut en soi, ça rend peut-être les choses claires, mais je dis, en termes de, je trouve que c'est assez classique. C'est pour ça que j'ai mis à peu près en dernier. Et c'est à peu près le lien avec le « *Novelty* », je suppose de juger, la façon dont ça s'est fait, j'ai trouvé que c'était assez classique. Bon, maintenant je sais bien que qu'on ne peut pas arriver [pif] conscient, pardon. Alors, après, c'était vraiment, je précise que c'était bien, mais [pause].
Interviewer	Donc vous les avez classés par rapport à une application santé, donc par rapport à ce que vous voudriez voir dans une application santé, ou par rapport à Babylon, spécifiquement ?
Participant 12	Je l'ai fait, par rapport à mon vécu, donc peut-être sur mes attentes de l'application santé de manière générale, en fait. Je pense que j'ai fait vis-à-vis de cette position-là.
Interviewer	D'accord.
Participant 12	Ensuite, en termes de « *Dependability* », c'est que, en fait, oui j'ai envie d'avoir confiance, mais je n'ai à peu près rien qui me permet de savoir si ce qu'on dit est bon. Ça m'avait l'air, ça avait l'air de l'être… « *Dependability* » je l'ai vraiment mis là parce que je ne savais pas trop où le mettre, en fait. Moi, ce qui m'avait le plus marqué, c'était vraiment l'aspect, l'attractivité, où là je trouvais que qu'on pouvait faire mieux. Après, ça a vraiment été, ça ne se jouait à rien. Donc, vraiment, l'attractivité, le « *Novelty* », je les ai mis, je ne sais pas si c'est sûr, mais après, j'ai vraiment modulé... ça s'est joué sur rien. « *Usefulness* », j'ai trouvé que ça l'était, ça l'était quand même. Et j'ai bien aimé le fait qu'il nous dise quand même d'aller voir un médecin, ou aller chez un pharmacien. Donc, moi, ce

que j'avais vu, j'avais vu qu'on peut aller chez un médecin on peut avoir les médicaments chez un pharmacien. J'ai trouvé ça utile, assez concret. Donc, je trouvais que c'était en lien avec le contenu et la qualité, donc, là je remonte, parce que déjà il y a le diagnostic qui a été fait, et même sur toutes les questions qui ont été posées, c'est assez complet. Je trouve qu'ils ont essayé de faire un lapse de panels de choses. Ce qui explique pourquoi je trouvais que le processus était lent. Tout à l'heure, j'ai mis que le processus était lent, dans « slow », je l'ai mis à 0, je l'ai mis neutre, parce que c'est un peu long, mais en même temps, si on veut avoir un diagnostic correct, je pense qu'il faut le faire. Dans [pi] j'ai quand même [pi], mais bon, après je ne suis pas médecin donc si ça se trouve c'est [pi] un truc complètement [pi], mais pour moi, voilà, c'est assez efficace. Et donc, aussi, après, là on remonte, pourquoi, j'ai mis ça en deuxième [Trustworthiness of Content] parce que, il y avait, toujours cette sorte de bienveillance, où dès le début ou nous dit bien « on n'est pas des médecins, on n'est pas des médecins, allez consulter » machin, si c'est grave, ils n'oubliaient pas de le rappeler. Et donc, je pars de ce postulat-là, que s'ils me proposent cette donnée, c'est que ça doit être un minimum fiable. Voilà, ils ne font pas, ils ont aucune mauvaise intention, j'ai l'impression. Et justement, j'ai mis, je ne sais pas comment mettre en français, parce que, c'est peut-être perspicuité, mais ça on ne sait pas, enfin, je ne sais pas, mais je l'ai mis là parce qu'en fait, ce n'est peut-être pas le plus attractif, c'est juste que j'ai trouvé ça clair, personnellement. Je trouvais ça facile à comprendre, très simple. Enfin, ce qui fait que pour n'importe qui, qu'il soit jeune ou moins jeune, donc on va dire les technophiles et les technophobes, ils vont facilement s'y retrouver. C'est pour ça que j'ai mis ça en *number one*, parce que pour moi c'est la partie [pif] en fait. Voilà.

Interviewer OK. Est-ce que le produit correspondait à vos attentes ?

Participant 12 Est-ce que j'avais des attentes ? C'est ça que... Ouais, est-ce que j'avais vraiment des attentes ? Bizarrement, si moi je devais faire une application comme ça, moi aussi j'utiliserais un peu ce, enfin, cette même façon de faire en, comme ça, boom ! Quelqu'un dirait c'est moins attractif [rire] mais, je pense que c'est ce que j'utiliserais pour toucher un maximum de personnes. Donc, oui, je pense que c'est ce à quoi je m'attendais. Ouais, je pense que, c'était, ouais. Si je devais répondre, je dirais que oui, quand même. Je dirais beaucoup oui, que non, parce que voilà. Je ne sais pas si vous voulez que je développe plus, mais.

Interviewer Non, c'est bien.

Participant 12 OK.

Interviewer Qu'est-ce que vous avez le plus apprécié dans le produit ?

Participant 12 C'est les résultats. Je trouvais que les résultats étaient assez complets, j'ai même cliqué sur tous les items, pour voir, etc. Je trouvais que c'était assez complet. Encore une fois, ils rappelaient, ils disaient ce qu'on « devait faire », soit voir un médecin, soit aller chez le pharmacien. Je trouvais ça assez bien. Après, je trouvais que c'était documenté. Les résultats étaient assez documentés. Je n'avais pas l'impression d'être face à des « imbéciles », [rire] à des charlatans, on va dire comme ça. Moi je dirais, c'est vraiment les résultats qui m'ont un peu étonné. Pour tout ce qui est vraiment [pif] ce qui fallait choisir, je n'étais pas, en fait, je n'étais pas surpris, parce que c'est comme ça que j'aurais fait. Vu que moi j'aurais peut-être fait ça comme ça, et que peut-être j'ai déjà vu ça dans d'autres types de plateformes et d'applications, [pif] je n'étais pas surpris, mais je trouve que ce n'était pas un tort en soi, vu que je trouve que ça reste facile à utiliser. Donc ouais, les résultats. Les résultats vraiment documentés et la possibilité, du coup, de prendre rendez-vous. C'est pour ça que j'étais curieux, en fait. J'étais curieux de voir ce que ça faisait, en fait. Donc voilà. J'ai trouvé qu'il y avait une action qui disait le diagnostic.

Interviewer OK. Qu'est-ce que vous avez le moins apprécié dans ce produit ?

Participant 12 Je dois mettre de côté l'anglais ? Parce que ouais, qu'est-ce que j'ai moins apprécié ? Je ne peux pas dire que j'ai passé un mauvais moment, en fait. Je ne sais pas, mais, c'est peut-être lié à l'interface, mais j'ai trouvé assez neutre, en fait. Je trouvais assez neutre, même peut-être visuellement, ce n'est vraiment pas très d'attractif, quoi. C'est, oui, toujours le même truc quoi, voilà c'est le truc peu attractif. Parce que moi, j'aime bien, moi, je suis le genre de gars, quand il fait une synthèse, il met des blagues dedans, vous voyez ? Pour que quand je les étudie, je passe un bon moment et mes potes aussi passent un bon moment [pi] blagues dedans. Donc, ouais, moi, c'est peut-être lié à ma personnalité, mais moi j'ai besoin de visuels, de choses accrocheuses, que ça soit des blagues ou pas, mais des choses accrocheuses ou... C'est vraiment ça, parce que la prise en main, je l'ai trouvée assez intuitive, ce n'est pas très compliqué. C'est [pi] la majorité du temps, où on a mal, ça fonctionne. Donc, ce que j'ai moins aimé, ouais l'attractivité, peut-être. La police ne m'a pas dérangé. Je ne sais pas si c'est un critère, enfin, je sais bien que ça

rentre dans [pi] mais, au final, la police ne m'a pas dérangé. Elle ne m'a pas marqué, en tout cas, vu que voilà. Les couleurs non plus. Voilà, c'était une espèce de mauve-pourpre. Pareil, pas de problème, mais ça reste assez classique. Donc voilà. Si je devais dire, pas attractif, peut-être, voilà.

Interviewer OK. Est-ce que vous aviez déjà utilisé une application similaire auparavant ?

Participant 12 Je ne sais pas comment expliquer, parce que j'arrêtais de me faire la réflexion, mais je ne voulais pas trop parler, parce que je parle tout le temps, mais en fait, je ne sais pas si vous connaissez, je crois que ça s'appelle Akinator. Vous connaissez Akinator, ou pas du tout ? C'était un truc hyper populaire, ce n'est pas la même chose, mais ça fonctionne à peu près pareil, mais c'est en gros, on veut faire deviner une star à Akinator et il doit poser des questions « est-ce que c'est un homme ou une femme », et puis, il va demander l'âge, la profession aussi, etc. Et, en gros, l'objectif c'est de pas de pouvoir utiliser Akinator, mais il gagne tout le temps parce que c'est l'algorithme qui [pi] donc ça, on est à mille lieues, parce que là justement il y a des petits dessins et tout on est à mille lieues de ça. Donc, ça c'est vraiment un truc beaucoup moins sérieux, on est vraiment dans le truc beaucoup moins sérieux. Est-ce que j'en avais déjà utilisé ? Non, en tout cas, non. Aussi longtemps, je ne pense pas. C'était peut-être pour deux, trois questions, comme ça, mais de mémoire, non, je ne pense pas. Je ne pense pas.

Interviewer OK. Hm. Je vais vous laisser terminer tranquillement le questionnaire. Vous pouvez arrêter le partage.

Participant 12 Ouais.

Participant 13

Niveau d'anglais annoncé : C1

Trajet dans l'application : idéal

Maitrise de la langue : excellent

Commentaire : /

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ?
Participant 13	Oui.
Interviewer	Pourquoi est-ce que vous les avez hiérarchisés de cette manière ?
Participant 13	Parce que je pense, je suis même sure, que pour une application, la confiance qu'on peut accorder aux résultats est vachement plus importante que le design de l'interface en elle-même. Parce qu'on peut me mettre une interface qui n'est pas si attirante que ça, que si les résultats que j'ai à la fin sont vraiment très clairs et auxquels je peux faire confiance, je préfère largement ça à une interface qui m'attirait peut-être mais dont je ne suis pas vraiment sûr de l'importance que je peux accorder aux résultats.
Interviewer	D'accord. Comment décririez-vous votre expérience de Babylon Health ?
Participant 13	Je trouvais que ça réunissait les deux, parce qu'il y avait énormément de questions qui étaient posées. On ne posait pas juste quelques questions en rapport avec les symptômes et puis on balance juste un détail comme ça, je trouve que c'était quand même très détaillé au niveau des questions par rapport aux symptômes qu'on a. Bon, voilà, personnellement je connaissais pas du tout les deux infections qui m'étaient proposées à la fin, je pense que c'est quand même, en tout cas moi j'ai quelque chose d'assez positif qui est sorti de cette interface.
Interviewer	Est-ce que le produit correspondait à vos attentes ?
Participant 13	Je n'avais pas vraiment d'attentes au niveau du produit, donc je ne sais pas répondre à cette question. Mais en tout cas c'est positif.
Interviewer	[rire] D'accord. Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 13	Justement, le fait qu'il y avait énormément de questions pour vraiment rechercher au détail près tous les symptômes que le patient a.
Interviewer	Et qu'est-ce que vous avez le moins apprécié ?
Participant 13	Je dirais parfois je ne sais pas réellement, en fonction des symptômes qui m'étaient présentés, et en fonction de ce que l'application me demandait, parfois ce n'est pas exactement le symptôme que l'on a, Donc on ne

	sait pas vraiment si on doit mettre oui, non, ou je ne sais pas, ou un symptôme qui y ressemble
Interviewer	Hm-hm. Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 13	Non, non. Pas vraiment.
Interviewer	Est-ce que vous aviez déjà utilisé une application similaire auparavant ?
Participant 13	Non, jamais.
Interviewer	OK. Je vais vous laisser terminer le *survey* tranquillement. Vous pouvez arrêter le partage.
Participant 13	D'accord.

Participant 14

Niveau d'anglais annoncé : « ok »

Maitrise de la langue : Bonne

Trajet dans l'application : non idéal

Commentaire : le participant est parti durant l'expérience. L'expérience a donc duré plus d'une heure.

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ?
Participant 14	Je dois revenir en arrière ?
Interviewer	Oui. Est-ce que vous pouvez revenir en arrière ?
Participant 14	Hm.
Interviewer	J'aurais quelques questions à vous poser. Vous préférez anglais ou français ?
Participant 14	Français.
Interviewer	D'accord. J'aimerais savoir pourquoi est-ce que ces qualités sont importantes pour vous, pourquoi est-ce que vous les avez hiérarchisées de cette manière.
Participant 14	Parce que je trouve qu'une application de santé, le plus important, ce n'est pas son interface, c'est son contenu qui compte, parce que ceux qui cherchent ce genre de formes c'est pour des informations utiles, des informations qui peuvent aider vraiment. Et la nouveauté ou l'attraction du site est secondaire je pense.
Interviewer	D'accord
Participant 14	Hm. Et le plus important est la faisabilité et la praticité.
Interviewer	OK. Comment décririez-vous votre expérience de Babylon Health ?
Participant 14	Ca va, ce n'est pas très difficile de prendre en main.
Interviewer	Est-ce que le produit correspondait à vos attentes ?
Participant 14	Oui, je pense que c'est pas mal, le produit.
Interviewer	Qu'est-ce que vous avez le plus apprécié dans le produit ?
Participant 14	Le plus... J'apprécie le plus les choix qu'elle me donne.
Interviewer	OK. Et qu'est-ce que vous avez le moins apprécié ?
Participant 14	Le moins apprécié c'est peut-être, parfois je n'ai pas trop compris les explications des symptômes.
Interviewer	Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 14	Je pense que non.
Interviewer	Est-ce que vous aviez une application similaire auparavant ?
Participant 14	Oui, il me semble que je le vois quelque part, mais ce n'est pas le même type d'application.
Interviewer	D'accord
Participant 14	Mais la logique est la même.

Interviewer	OK. Super, je vais vous laisser terminer le *survey* tranquillement. Vous pouvez arrêter le partage.
Participant 14	D'accord, merci. [quitte la conversation]

Participant 15

Niveau d'anglais annoncé : B1

Maitrise de la langue : Suffisant

Trajet dans l'application : il a d'abord fait covid-check. Trajet dans l'application OK

Commentaire : J'étais en retard de 5 minutes car l'expérience précédant celle-ci a duré plus d'une heure. J'ai prévenu le participant 15 de mon retard.

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière, s'il vous plait ?
Participant 15	Ouais.
Interviewer	Pourquoi est-ce que vous les avez hiérarchisés de cette manière ? Pourquoi est-ce que ces qualités sont importantes pour vous ?
Participant 15	Le premier objectif c'est d'être efficace et utile. Si on va sur une application de santé, pour moi, c'est d'abord l'objectif final qui est de décrire nos symptômes, enfin en tout cas, avoir l'idée la plus claire possible sur ce qu'on a. Donc, l'objectif final. Et puis alors, du coup, la réponse qu'on a, et le résultat, faut que ça soit fiable. Après seulement, pour moi, vient l'attractivité, le design de l'application. En tout cas pour une application santé, ça vient après coup, même si c'est important aussi parce qu'il faut qu'on puisse s'y retrouver et il faut que ça soit assez éclairé au niveau design. Mais, pour moi, c'est d'abord le produit final, le résultat final qui est important.
Interviewer	Comment décririez-vous votre expérience de Babylon Health ?
Participant 15	Pas facile à se faire comprendre au début. Après, ça vient petit à petit. Mais les questions viennent un peu dans un ordre, je ne dirais pas aléatoire, mais un peu genre, au début, je parle plutôt de mon oreille, etc. Et après, je parle de ma mâchoire. Enfin, voilà, ce n'est pas facile à décrire, mais, au final, je pense que le résultat était satisfaisant pour moi.
Interviewer	Est-ce que le produit correspondait à vos attentes ?
Participant 15	Oui, relativement, oui.
Interviewer	Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 15	Le fait que ça soit des questions, etc., on va dire fermées, ce qui fait qu'on peut répondre des trucs concrets. Bon, a part quand on peut répondre « je ne sais pas ». Je pense que ça, c'est pour être le plus

	efficace possible et que le processus voit directement où nous diriger.
Interviewer	Et qu'est-ce que vous avez le moins apprécié ?
Participant 15	Je vais dire, plus ou moins, l'ordre des questions dans lesquelles elles venaient. Parfois, ça perturbait un petit peu ce qu'on était, la partie, le symptôme auquel on était en train de répondre. Voilà, ça c'était un peu plus difficile.
Interviewer	Donc, est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 15	Non, pas vraiment. À part pour la partie au début, où j'ai répondu au covid-check. Enfin, dans la tête c'était plus une formalité qu'autre chose, mais visiblement, l'application voyait ça un peu comme un passe-droit. Si on répondait de manière un peu inquiétante, on était déjà catégorisés, du coup bah, voilà. Alors que, dans ma tête, je pensais que c'était plus une formalité pour commencer le questionnaire.
Interviewer	D'accord. Est-ce que vous aviez utilisé une application similaire auparavant ?
Participant 15	Non.
Interviewer	Super. Je vais vous laisser continuer de répondre aux questions tranquillement. Vous pouvez arrêter le partage.

Participant 16

Niveau d'anglais annoncé : C2

Maitrise de la langue : Suffisant (loin d'un C2)

Trajet dans l'application : pas idéal, résultats rien à voir

Commentaire : À la demande du participant, j'ai traduit l'entièreté de la situation. Il a souvent utilisé le bouton « *I don't know* ». Le participant a inventé des symptômes. Il s'est fait interrompre durant l'expérience.

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ? [pause] Pourquoi est-ce que vous avez hiérarchisé les qualités de cette manière ?
Participant 16	Là, en premier, j'ai mis ce que je trouvais le plus important.
Interviewer	Oui. Pourquoi c'est important pour vous ?
Participant 16	Moi, je trouve qu'avoir des informations fiables soit le plus important. Je n'aimerais pas qu'on me donne des informations qui ne sont pas fiables, surtout pour le domaine de la santé. Ensuite, que ça soit efficace, c'est un peu pareil, je pense. Faut aussi de la qualité. Ensuite c'est important que ça soit clair pour tout le monde. Après, moi, l'attractivité, je ne trouve pas ça très important. Je trouve qu'il n'y a pas besoin d'avoir de l'attractivité puisqu'on utilise ça uniquement quand on est malade. Donc, si quelqu'un est malade, il l'utilisera, si quelqu'un n'est pas malade, il ne l'utilisera pas. Et « *Novelty* », c'est dans le sens dans le sens où c'est nouveau, c'est plus ça ?
Interviewer	[pi]
Participant 16	Je ne trouve pas que c'est important. Enfin, pour moi ce n'est pas un des critères les plus importants
Interviewer	Comment décririez-vous votre expérience de Babylon Health ?
Participant 16	Intéressante. Ouais, intéressant. C'est une bonne idée je trouve, pour se faire un premier avis. Parce que moi je vais voir, personnellement, quand j'ai des symptômes sur internet. Plus par peur de voir, enfin, des trucs faux, de m'inquiéter pour rien. C'est pour ça j'aimerais bien un truc fiable, avec des bonnes informations
Interviewer	Est-ce que le produit correspondait à vos attentes
Participant 16	Oui ?
Interviewer	Est-ce que vous aviez des attentes ?
Participant 16	Non, je n'avais pas réellement des attentes. Donc non.
Interviewer	Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 16	Le fait que... Je ne sais pas comment dire. Ça m'avait l'air bien précis. Bon, après, moi je ne parle pas très

	bien anglais donc ça m'a peut-être plus posé problème. Mais sinon, ça avait l'air d'être bien, les informations avait l'air d'être pas mal.
Interviewer	Qu'est-ce que vous a le moins plu ?
Participant 16	Peut-être le fait quand ça soit en anglais, justement, parce que justement je n'ai pas tout bien saisi. Après, je sais que j'aurais pu vous demander, mais je vois, je ne voulais peut-être pas vous déranger.
Interviewer	Vous ne me dérangez pas du tout, je suis là pour ça. Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 16	Non, non.
Interviewer	Est-ce que vous aviez utilisé une application similaire auparavant ?
Participant 16	Jamais, non.
Interviewer	D'accord. Je vais vous laisser terminer de remplir les questionnaires tranquillement. Vous pouvez arrêter le partage.

Participant 17

Niveau d'anglais annoncé : « Moyen »

Maitrise de la langue : Suffisant

Trajet dans l'application : pas idéal (a commencé par parler du nez), n'a pas trouvé des résultats adéquats

Commentaire : Participant 17 a eu de nombreux problèmes techniques et des difficultés à se connecter à Zoom. Le Limesurvey ne fonctionnait pas durant le premier rendez-vous, donc nous avons dû reprogrammer au jour suivant.

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ?
Participant 17	Je fais juste ... ?
Interviewer	Pourquoi est-ce que vous les avez hiérarchisés de cette manière ?
Participant 17	Parce que... Je ne sais pas, c'est comme ça, c'est mon avis, voilà.
Interviewer	D'accord. Pourquoi est-ce que ces qualités sont importantes pour vous ?
Participant 17	Parce que c'est un mixte entre quelque chose d'efficace, et quelque chose de confiance, il faut que ça le soit d'office. Je ne sais pas, c'est dans mon inconscient, je vois ça comme ça.
Interviewer	Comment décririez-vous votre expérience de Babylon Health ?
Participant 17	Pardon ?
Interviewer	Comment décririez-vous votre expérience de Babylon Health ?
Participant 17	C'est, le design, tout ça, c'est quelque chose d'agréable, [pi] mais, je pense que quelque chose qui doit être aussi précis, comme les symptômes de la médecine, c'est difficile de, par exemple, je vous ai posé la question, si je pouvais rentrer un seul symptôme au début, bah, ça a fait toute la continuité là-dessus, mais il en avait d'autres choses qui étaient plus importantes. Par exemple, on parlait de l'oreille, de la douleur à l'oreille, et on parlait que c'était très important ça, c'était très douloureux, et je n'ai pas su mettre ça, il n'y a rien qui parlait de l'oreille plus tard. Donc, à moins de recommencer, plusieurs fois l'expérience, c'est un peu... Personnellement, je n'irais pas sur un site pour voir mes symptômes. J'irais voir directement un médecin, surtout si c'est la douleur que la personne exprimait. Moi je trouve ça à moitié efficace, mais, si ça peut rassurer des gens... C'est comme, Doctissimo ? C'est ça ? Ouais, c'était ça, Doctissimo, les gens regardaient, ils avaient mal au doigt, ils avaient le cancer du cerveau. Voilà, moi ce n'est pas ce que je préfère, ce n'est pas ce

	que je ferais, mais voilà. Mais il faut que ça soit le plus efficace possible, digne de confiance et ça, ça dépend de l'avis de chacun. Voilà
Interviewer	Est-ce que le produit correspondait à vos attentes ?
Participant 17	Non, je n'avais pas beaucoup d'attentes. Un peu, oui, quand même, parce que ça donne une petite description de ce qu'on pourrait avoir donc. Pas tout à fait, mais un peu, oui, Pas complètement, voilà.
Interviewer	Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 17	La facilité, page après page. C'est assez simple d'utilisation. Oui, la facilité. C'est ce qui me reste le plus en tête.
Interviewer	Et ce que vous avez le moins apprécié ?
Participant 17	Le fait de, justement, je trouve qu'il n'est pas complet. Le côté que ça ne soit pas [hésite] complet, tout simplement. Si on avait été chez médecin, ça aurait été différent, un échange différent. C'est peut-être plus facile aussi d'exprimer directement avec une personne qu'avec une intelligence artificielle, juste en répondant à des questions. Comme on a déjà des, c'est un questionnaire fermé, on choisit, on choisit, mais on ne peut pas non plus s'exprimer complètement.
Interviewer	Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 17	Je ne saurais pas vous dire.
Interviewer	Est-ce que vous aviez déjà utilisé une application similaire auparavant ?
Participant 17	Dans ce domaine-là ? De médecine ?
Interviewer	De manière générale.
Participant 17	De manière générale ? Je pense que oui, mais je n'en ai pas une qui vient en tête, en tout cas. Mais surement, en tout cas, oui.
Interviewer	D'accord. Je vais vous laisser terminer le questionnaire tranquillement. Vous pouvez arrêter le partage.
Participant 17	Oui, c'est vous qui décidez. Voilà.

Participant 18

Niveau d'anglais annoncé : B2

Maitrise de la langue : Bonne

Trajet dans l'application : pas idéal, a commencé avec « jaw »

Commentaire : le participant était en retard pour l'expérience, a essayé d'enregistrer l'entretien et a demandé une copie de l'entretien. Le participant s'est d'abord connecté avec sa tablette pour prendre des notes. La connexion internet était lente et instable. Le participant a inventé des symptômes.

Transcription de l'entretien

[Depuis la fin du 1er call, pas enregistré pendant +/-15min]
Résumé des minutes manquantes
Réponse à la 1ere question « Pourquoi est-ce que vous les avez hiérarchisés de cette manière ? / Pourquoi est-ce que ces qualités sont importantes pour vous ? »
Ranking 1 : Usefulness
Le plus important c'est que ça soit utile. Elle s'en fiche qu'il y ait déjà plusieurs produits similaires, le plus important c'est que ça marche et que le produit réalise les objectifs de son utilisation
Ranking 2 et 3 : Dependability (2) et Content Quality (3)
Ces deux qualités sont à valeur égale. Le plus important c'est de savoir qu'on peut faire confiance au produit, qu'il y a des médecins (des professionnels) derrière la création de l'outil et c'est sur base de leurs connaissances que le produit crée des diagnostiques. Faire confiance aux avis émis par le produit est primordial.
[suite transcrite grâce à l'enregistrement]

Participant 18 Une personne qui a peut-être plus de compétences et de connaissances, que quelqu'un qui n'en a pas, ou qui en a moins. Enfin, il faudrait qu'un large public puisse comprendre, donc la clarté, et que ça doit être efficace. Que ça m'aide, que ça m'aide, que ça m'aide réellement, que je puise, voilà. Le fait que ça soit attrayant, [pf]. Je préfère plus que ça soit efficace ou utile, qu'attrayant. Donc, voilà. Et du coup, la nouveauté, ça reprend, ça prend le même, c'est comme, un peu, [l'attirance], c'est que je préfère que ça soit un truc revu qui a été retravaillé et qui est devenu bien et qui est devenu de qualité, plutôt que ça soit quelque chose de nouveau. On est attiré justement par justement, le fait que ça soit nouveau, attractif, mais que, dans le fond, le contenu, l'utilité, et tout ça ne soit pas pris en compte. Donc voilà.

Interviewer Ok. Et comment vous décririez votre expérience de Babylon Health ?

Participant 18 Personnellement, au début, je n'avais pas compris qu'au début, je n'avais pas compris que c'était un symptôme. Donc, voilà. Après... C'est pour ça qu'à un moment j'ai répondu que ce n'était pas trop compréhensible, dans le sens ou si on vient et qu'on ne connait pas, eh bah, on fait des tests, et du coup on clique plusieurs fois, on clique beaucoup trop. Moi je suis qu'un qui essaie de limiter mes clics, et donc, voilà. Si je dois essayer de comprendre, et puis après, enfin, voilà. C'était un peu incompréhensible au début. Après, quand on comprend, je comprends le but, mais du coup, il y a eu, un moment, il y a eu une question « est-ce que c'est lent ? ». Du coup, je trouve que c'est lent, que c'était un peu plus lent. Et par rapport à la réponse que j'ai eu, par rapport aux symptômes que j'ai dû mettre [hésite] je n'ai pas eu l'impression qu'on me répondait réellement. J'ai eu l'impression, qu'on m'a dit « voilà, vous avez répondu à tout ça, ça peut être grave, donc allez voir un médecin ». J'ai eu l'impression que c'était plus ça. Je n'ai pas eu de piste de ce que j'aurais pu avoir, de petites idées, je n'ai pas eu tout ça. Voilà. J'ai juste eu « ça peut être grave donc allez voir un médecin ». Mais vu qu'à la base des bases, avant de faire de remplir mes symptômes, ils me disent déjà que de base, voilà, ce n'est pas exhaustif, qu'il faut aller voir un médecin, que c'est toujours préférable, bah, me dire à la fin qu'il faut aller voir un médecin, sans me donner de pistes, eh bah, je n'ai pas l'impression alors d'avoir, que ça a été très, très utile de faire ça. Si c'était pour faire ça, alors autant aller directement chez le médecin. Sinon, ce n'est pas très, c'est un bon dispositif, je trouve. C'est un bon dispositif, et avec la société, comment elle a évolué, comment elle est à distance, je trouve que ça peut vraiment être bien. Ça peut vraiment être bien et ça peut vraiment remplacer les recherches google qui ne servent strictement à rien. Donc, l'idée, je la trouve bien. Mais après, je t'avoue que je n'étais pas trop satisfaite. Voilà.

Interviewer Donc, est-ce que le produit correspondait à vos attentes ?

Participant 18 [hésite] je n'avais pas trop d'attentes donc [expire] je ne saurais pas dire ça, mais je ne peux pas... Je dirais, vous voyez, quand sur votre échelle, il y avait « 0 » et « + », et puis « ++ » ? Je serais entre le « + » et « ++ ».

Interviewer Ok.

Participant 18 Si je pouvais exprimer ça comme ça, je ne sais pas. Ouais.

Interviewer	Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 18	Dans ce projet ?
Interviewer	Dans ce produit.
Participant 18	Dans ce produit. Moi, j'aime bien le produit en général. Comme je vous ai dit, le produit en général, je trouve que c'est une super, c'est une super bonne idée. Le produit, en tant que dispositif, je l'aime bien. Après, comme je l'ai dit, je trouve ça un peu lent et je ne trouve pas ça assez clair, vers la fin. Donc, voilà.
Interviewer	Et donc, qu'est-ce que vous avez le moins apprécié dans ce produit ?
Participant 18	J'ai moins apprécié le diagnostic. Je dirais le diagnostic. J'ai moins apprécié le diagnostic parce que, je ne sais pas si c'est par rapport aux symptômes, ou si c'est par rapport en règle générale. Je me suis demandé aussi si, si j'avais fait avec des symptômes, par exemple, pour le rhume, est-ce qu'il m'aurait dit que j'avais un rhume ? Je me suis posé la question de savoir ça. Donc, je ne sais pas si c'est réellement par rapport aux symptômes parce que les symptômes choisis font que c'est la seule réponse qu'ils aient pu me donner. Mais, si c'est le genre de réponse, alors c'est la réponse qui m'a le plus déçue, parce qu'alors je n'ai pas l'impression qu'un m'ait répondu. J'ai l'impression que m'on m'a redit ce qu'on m'a dit au début, je n'ai pas l'impression qu'on m'ait réellement répondu.
Interviewer	Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 18	Non. À part mon ordinateur [rire]
Interviewer	Est-ce que vous aviez déjà utilisé une application similaire auparavant ?
Participant 18	Non.
Interviewer	Super. Je vais vous laisser terminer tranquillement le questionnaire. Vous pouvez arrêter le partage.
Participant 18	Ok, d'accord.

Participant 19

Niveau d'anglais annoncé : « intermédiaire »

Maitrise de la langue : Suffisante

Trajet dans l'application : A réalisé un covid check. Trajet pas idéal, car il a commencé avec des symptômes relatifs au nez.

Commentaire : l'expérience a commencé en retard car le participant ne s'est pas rendu compte que c'était sur Zoom. Participant 19 a rapidement lu le formulaire de consentement au début de l'expérience.

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ?
Participant 19	D'accord. Ah ! Allez, non. Oh, putain, qu'est-ce que j'ai foutu. D'accord ouais. Oui.
Interviewer	Est-ce que vous pouvez m'expliquer pourquoi est-ce que ces qualités sont importantes pour vous ? Pourquoi est-ce que vous les avez hiérarchisées de cette manière ?
Participant 19	D'accord. Eh bien, euh, j'ai trouvé que le fait que le produit soit efficace et que « *Perspicuity* » c'était bien perspicacité, je pense.
Interviewer	*Perspecuity*, si vous remontez un petit peu...
Participant 19	Je n'ai pas compris.
Interviewer	Vous pouvez remonter un tout petit peu ? Sur la page. Voilà. *Perspicuity* c'est donc la qualité d'être clair et facile à comprendre.
Participant 19	Ouais c'est ça, Donc, en soi, oui, parce que j'ai trouvé ça efficace. Bon, c'est quand même assez facile à comprendre quand même. Pour moi, le fait que l'efficacité soit en premier lieu étant... est en première ligne, dans le sens où il faut que le produit soit bon pour qu'il soit sur le marché, en vrai. Donc, je me suis dit que ça, c'était quand même quelque chose de sensible, surtout à l'heure actuelle où tout pousse - où tout passe par la technologie et donc... Le deuxième, c'était facile à comprendre. Ben justement, il faut que ça soit en symbiose avec l'efficacité et la qualité de contenu. Pour moi c'est... En tout cas c'est dans mon top 3, dans le sens où ça fait partie d'un tout, quoi. Ça fait partie d'un tout, et je pense que tout doit être, sera top 3 équitable. Et il faut que [pause] allez, la qualité du contenu soit... soit aussi dans ce top 3. Parce que sinon, si on ne relaye pas un contenu de qualité, le produit est un peu désuet, en quelque sorte, quoi. Donc voilà, c'était un peu mon objectif et évidemment, il faut qu'il y ait quand même un certain, un certain aspect de nouveauté, dans le sens où aujourd'hui, on connaît beaucoup les applications et par rapport au coronavirus

et aux symptômes par rapport à d'autres maladies, bah c'est quand même quelque chose de super actuel et pour moi, bah, je trouvais ça très intéressant de mettre ça dans mon top 4.

Interviewer OK. Donc, ça, c'est où les qualités les plus importantes dans une application santé de manager générale ?

Participant 19 Hm-hm.

Interviewer Super. Et comment décririez-vous votre expérience de Babylon Health ?

Participant 19 Eh bien, par rapport à mon expérience en tout cas, le fait que ça a été en anglais m'a demandé plus de réflexion. Bon, maintenant, c'est un très bon exercice aussi pour moi. Sinon, je pense que mon expérience personnelle avec cette application fut quand même quelque chose de bénéfique dans le sens où je vois qu'aujourd'hui plus que jamais, bah c'est important de promouvoir ce genre d'applications dans le sens où aujourd'hui, on est très confronté à ce genre de virus et ça nous permet aussi de faire la part des choses entre les virus et les symptômes d'une autre maladie, par exemple. Donc qui peuvent nous dire si, par exemple, par rapport à ce que nous avons comme symptômes, bah est ce qu'on a le coronavirus ? Oui ou non. Et c'est quand même très essentiel. Aujourd'hui, je trouve.

Interviewer OK. Qu'est-ce que vous avez le plus apprécié dans ce produit ?

Participant 19 Le plus apprécié... Ben déjà, c'est ce que je viens de vous dire par rapport à ce concept de coronavirus qui est très actuel. Donc ça, c'est bien que le fait que cette application reprenne tout ça sous une forme d'application, tout simplement. C'est peut-être un peu bête à dire, mais c'est facile à utiliser et comme je l'ai dit, c'est efficace. Donc effectivement, ça m'a plu dans le sens où c'est quand même très intuitif et ça peut aider pas mal de personnes qui ne connaissent pas le... Je vais dire, la technologie d'aujourd'hui, comme ma maman qui, elle, pourrait l'utiliser facilement si elle comprenait l'anglais, quoi. Et ça peut permettre justement à des personnes plus âgées qui n'ont pas cette facilité d'utilisation de ce genre d'applications, bah à l'utiliser plus facilement.

Interviewer OK. Et qu'est-ce que vous avez le moins apprécié ?

Participant 19 Le moins apprécié, c'est toujours la question la plus difficile... Mais le moins apprécié, euh... je dirais qu'il faut répondre spontanément que pour moi, il n'y a pas quelque chose qui me semble à retravailler, sauf... Le côté attractif pour moi, il était bon. Non, je préfère rester spontané et me dire qu'il ne faut pas forcément retravailler un truc qui est toujours bon, sauf euh - enfin, qui est bon dans ce cas-ci - sauf tout à l'heure. Je

	me souviens que j'ai mis dans mon choix de 0 à 10, enfin, si je peux le dire comme ça, quand c'était très [inaudible] ou pas, j'avais mis certains certaines choses à neuf à la place de dix, dans le sens où, pour moi, un produit nécessite toujours d'être, d'être retravaillé. Il nécessite toujours un suivi. Mais bon, voilà, je n'ai pas forcément envie de mettre ça dans un point négatif, quoi. Mais c'est une réponse que j'ai envie de donner spontanément.
Interviewer	D'accord. Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 19	De manière inattendue dans le sens où je n'ai peut-être pas très bien compris le début, où c'était un cas de covid ou non. Et c'est le seul inattendu que j'ai reçu. Par rapport à l'application, je pense que tout, tout est fluide. Tout a bien, tout a bien suivi les bonnes étapes et je pense que je me suis bien retrouvée dans le programme. Donc non, pas vraiment d'inattendu, sauf les huit réponses ici sur le dernier. Le dernier truc à faire là, c'était peut-être inattendu, dans le sens où il ne pas forcément compris qu'il fallait mettre les 8 dans un *ranking*, tout simplement. Je pensais qu'il fallait mettre un maximum de 8 et ça, je pense que c'est plus par rapport à ma compréhension de l'anglais.
Interviewer	OK. Est-ce que vous avez déjà utilisé une application similaire auparavant ?
Participant 19	Non. Enfin, par rapport au covid, non. Euh... similaire à celle-ci, honnêtement... Je suis sûr que si je réfléchis, je pourrais en trouver une, mais comme ça, à froid, je ne pense pas, non.
Interviewer	OK. Je vous laisser terminer le questionnaire tranquillement- Vous pouvez arrêter le partage.

Participant 20

Niveau d'anglais annoncé : B2

Maitrise de la langue : excellent

Trajet dans l'application : bon trajet dans l'application

Commentaire : la participante était malade la première fois, nous avons donc reprogrammé l'entretien pour la semaine suivante.

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ?
Participant 20	Oui, pas de soucis.
Interviewer	Pourquoi est-ce que vous avez hiérarchisé ces qualités de cette manière ? Pourquoi est-ce qu'elles sont importantes pour vous ?
Participant 20	Parce que pour moi, avant que ça soit attractif, il faudrait que ça soit utile. C'est pour ça que j'ai commencé par ces deux-là, dans ma tête, peut-être pas dans le placement des choses. Et après, j'ai fait vraiment par ordre ce que je t'attendrais face à un médecin, en fait. J'attendrais que je puisse lui faire confiance, qu'il soit efficace, et pas qu'il soit joli à regarder. Du coup, je l'ai vraiment pris, voilà, comme quelque chose de vraiment utile de professionnel, on va dire. Plutôt que quelque chose ou je peux juste poser mes questions sur un symptôme.
Interviewer	Comment décririez-vous votre expérience de Babylon Health ?
Participant 20	C'était chouette [rire]. Franchement, je ne connaissais pas, mais s'il y a l'équivalent en Belgique, par exemple, pour pouvoir appeler directement les médecins, franchement, je trouverais ça super sympa. C'était facile d'utilisation. Enfin, c'est assez intuitif, j'ai envie de dire. Il y a deux trois questions où j'aurais aimé aussi voir la mention « je n'en sais rien ». Parce que, parfois, soi-même, on ne sait pas quel genre de douleurs on a et devoir mettre un mot dessus, ou, ce n'est peut-être pas aussi simple. Mais, c'était facile.
Interviewer	Est-ce que le produit correspondait à vos attentes ?
Participant 20	Oui, oui. Oui, vraiment. Ça fait ce que c'est sensé faire [rire]
Interviewer	Hm-hm. Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 20	Je dirais que le fait, que quand on, allez, l'organisation logique des questions, en fait. Parce que j'ai déjà eu des tests comme ça, sur des plateformes qui se ressemblaient un peu, où les questions se ressemblaient un peu, ou les questions sont... Ù il y a une espère ce chemin logique dans les questions et ça ne se suivait pas forcément. Alors qu'ici, si on dit que j'ai mal à la

	gorge, on va continuer sur la gorge avant de passer à autre chose. Ou avec les oreilles, pareil, vraiment, d'avoir vraiment fait le tour de la chose, d'avoir, pardon, fait le tour du problème, avant de passer au symptôme suivant. Et ça, je trouvais vraiment intéressant.
Interviewer	Et ce que vous avez le moins apprécié ?
Participant 20	Je dirais juste le fait qu'il manquait, ouais, peut-être, le petit bouton « je ne sais pas » pour certains symptômes. À part ça, rien.
Interviewer	Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 20	[hésite] Ah, oui, oui. Le fait de pouvoir prendre rendez-vous avec le médecin. Je ne m'attendais pas à voir ça à la fin.
Interviewer	Et, donc, est-ce que vous avez déjà utilisé une application similaire auparavant ?
Participant 20	Pas pour la santé, mais j'ai déjà utilisé une application qui faisait la même chose, en termes de chemin logique, mais en question santé, non.
Interviewer	OK. Super. Je vais vous laisser continuer le *survey* tranquillement.
Participant 20	Pas de soucis.
Interviewer	Vous pouvez arrêter le partage.
Participant 20	OK. Voilà.

Participant 21

Niveau d'anglais annoncé : B2

Maitrise de la langue : Suffisant

Trajet dans l'application : bon

Commentaire : participant 22 était très mécontent et disait souvent ne pas comprendre. Participant 22 se plaignait que c'était difficile à comprendre et que ce n'était pas très bien expliqué. J'ai dû le coacher un peu au début. Il semblait un peu perdu lors de l'entretien.

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière ? Pourquoi est-ce que vous avez hiérarchisé les qualités de cette manière ? Pourquoi est-ce qu'elles sont importantes pour vous ?
Participant 21	« *Perspicuity* » là, il faut que ça soit clair et facile à comprendre, quand on est sur une application. Sinon, c'est, [pif]
Interviewer	C'est l'efficacité ?
Participant 21	Enfin, oui, que ça soit, ça serve à quelque chose. [pif]
Interviewer	Je vous ai mal entendue, désolée.
Participant 21	Je ne suis pas sûre en fait, si « *Efficiency* » de ce que ça veut dire.
Interviewer	C'est l'efficacité.
Participant 21	Ah, oui. Donc. C'est efficace. Je dois dire tout ?
Interviewer	Hm. Ce que je veux comprendre, c'est, pourquoi est-ce que pour vous, l'efficacité, par exemple, ça fait partie des choses les plus importantes.
Participant 21	Quand on est sur une application, je trouve qu'il faut que ça soit efficace, qu'on choisisse, qu'on choisisse ce dont on a besoin. Comment on dit, que ça soit, efficace et en même temps que ça utile et en même temps, comment dire, que ça aille bien, que l'application, comment on dit. [pi]
Interviewer	Je suis désolée, je vous entends très mal.
Participant 21	Bah, voilà, je ne sais pas si je dois encore dire des trucs.
Interviewer	D'accord. Comment est-ce que vous décririez votre expérience de Babylon Health ?
Participant 21	[hésite] Je pense que parfois avec l'anglais, ce n'était pas toujours facile. Mais souvent, quand il y a la possibilité de « je ne sais pas », j'ai plus vite coché « je ne sais pas » parce que je ne savais pas ce que je devais mettre, en fait. Je ne sais pas si c'était moi, ou si je devais imaginer un personnage, ou pas, alors j'ai mis souvent « je ne sais pas ». Ça n'a rien à avoir avec l'application, c'est vraiment pour soi. Donc, on peut dire ce dont on a besoin, mais comme moi, je ne savais pas vraiment, j'ai mis « je ne sais pas ».

Interviewer	OK. Est-ce que le produit correspondait à vos attentes ?
Participant 21	Je n'avais pas d'attentes ? [rire]
Interviewer	OK. Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 21	A la fin, quand ça nous propose plusieurs diagnostics, et qu'il nous explique ce que c'était.
Interviewer	Hm-hm. Et qu'est-ce que vous avez le moins apprécié ?
Participant 21	Il y avait peut-être des questions qui étaient parfois un peu similaires, où la réponse était parfois similaire, donc parfois je me demandais un peu si j'avais bien répondu avant, ou si je devais répondre autre chose, puisqu'il y avait les réponses [pi]
Interviewer	Est-ce que quelque chose a fonctionné de manière inattendue ?
Participant 21	Non.
Interviewer	Est-ce que vous avez utilisé une application similaire auparavant ?
Participant 21	Non, jamais.
Interviewer	OK. Super. Je vais vous laisser terminer le *survey* tranquillement. Vous pouvez arrêter le partage.
Participant 21	D'accord.

Participant 22

Niveau d'anglais annoncé : B2

Maitrise de la langue : Excellent

Trajet dans l'application : idéal

Commentaire : /

Transcription de l'entretien

Interviewer	Est-ce que tu peux retourner en arrière ?
Participant 22	Ouais.
Interviewer	Pourquoi est-ce que tu as hiérarchisé les qualités de cette manière ? Pourquoi est-ce qu'elles sont importantes pour toi ?
Participant 22	En règle générale ?
Interviewer	Pour les applications santé.
Participant 22	Ouais. Parce que, hm. Allez. C'est bien d'avoir quelque chose d'attractif, mais si ça donne une réponse qui est [pf], trop alarmante ou qui ne me satisfait pas, bah [pf] ça ne sert à rien quoi.
Interviewer	Hm-hm.
Participant 22	Voilà. Est-ce que je dois dire par rapport au truc ici spécialement ?
Interviewer	Non. Je comprends pourquoi, du coup.
Participant 22	OK.
Interviewer	Et comment est-ce que tu décrirais ton expérience de Babylon Health ?
Participant 22	Ici, enfin, allez. Je recherche souvent quand même mes symptômes sur internet si j'ai des symptômes ou ce genre de choses.
Interviewer	Hm-hm.
Participant 22	Et en général, enfin, si je sais par exemple si je suis malade de quelque chose, donc ici des allergies, je vais d'abord mettre ça en premier et donc le fait qu'au tout départ, il n'y a pas de... Il y ait simplement... Il n'y ait pas le choix de... Enfin, on ne puisse pas... ça ne prend pas en compte les allergies, alors que je pense que ça joue un grand rôle ? Je vais moins croire l'application.
Interviewer	OK.
Participant 22	Parce [s'arrête dans son élan]
Interviewer	Non, continue.
Participant 22	Parce qu'après j'ai eu l'impression d'avoir dû faire tout le test et tout ça alors que je pense que ça aurait pu être divisé par deux si on avait su dès le départ que j'avais des allergies, tu vois ?
Interviewer	Ouais, je vois.
Participant 22	OK. D'autres questions ?
Interviewer	Oui. Et qu'est-ce que tu as le plus apprécié dans ce produit ?

Participant 22 J'aime bien quand même que ça prenne en compte tout, allez, même des choses auxquelles on ne penserait pas. Donc ça fait vraiment un diagnostic complètement total. Et aussi à la fin, il y a différentes « solutions » ? pertinentes et il est aussi possible de voir ce qui. Enfin, tu sais, t'as les deux solutions et t'as tout ce qui est risques et tout ça. Et dans le deuxième, bah voilà, il était noté « les allergies », tu vois.

Interviewer Hm-hm

Participant 22 Donc voilà. Tu peux aussi, toi, enfin même si au début ce n'est pas pris en compte, à la fin je pouvais aussi dire « OK, c'est plus probable que ça soit le deuxième étant donné que là il y a les allergies qui rentrent en compte ».

Interviewer Ouais. Et qu'est-ce que tu as le moins apprécié ?

Participant 22 Je t'ai dit, le fait qu'au départ, ça ne soit pas complet. Enfin voilà, il y a des fois où même en allant sur internet, on sait déjà ce qu'on a. Enfin, voilà.

Interviewer D'accord. Et est-ce que t'avais déjà utilisé une application similaire auparavant ?

Participant 22 Non, quand je suis malade, je vais sur internet, genre Google.

Interviewer Ça va. Tu peux arrêter de partager ton écran et continuer le *survey* tranquillement.

Participant 23

Niveau d'anglais annoncé : B2

Maitrise de la langue : bonne

Trajet dans l'application : bon

Commentaire : /

Transcription de l'entretien

Interviewer	Est-ce que tu peux revenir en arrière ? Pourquoi est-ce que tu as hiérarchisé ces qualités de cette manière ? Pourquoi est-ce qu'elles sont importantes pour toi ? Je ne t'entends pas.
Participant 23	Pardon [rire] Alors, pour moi, je me suis beaucoup, les trois premières, disons, pour moi c'est qualité, en fait, de ce qui t'es donné, et typiquement dans les autres questions, il y a des moments où je n'ai pas mis « +++ » parce que je n'ai pas les compétences médicales, en fait, pour savoir, si le contenu en soi, il est reliable ou pas. Enfin, tu vois. Est-ce que, du coup, le contenu est bon ou pas. Donc, pour moi, vraiment, c'est ça le plus important dans ce genre d'applications. Et puis, ensuite, du coup, si ce n'est pas compréhensible, ça ne sera pas utilisé. Donc ça reste assez haut dans mon *ranking*. Et puis, ensuite, pour moi, l'utilité, ça rentre un peu dans le fait que, si, bah voilà, si c'est compréhensible, etc., et que c'est utile, à ce moment-là, ça reste super important. Et après, je pense que le reste, c'est beaucoup moins important. Ça reste une application qui va te permettre d'avoir une idée sur les symptômes, ou le diagnostic que tu pourrais avoir. Et peut-être que toi-même, du coup, en utilisant ce genre d'applications, ça peut influencer en fait, en quelque sorte, ce que tu as et du coup tu vas te persuader que tu as quelque chose, en quelque sorte.
Interviewer	Hm-hm.
Participant 23	Et pour moi, le contenu, du coup, il est extrêmement important. Et je ne sais pas, est-ce que du coup.... Enfin, je remets un petit peu en question le fait que, est-ce qu'un robot pourrait quand même remplacer, en quelque sortes, les compétences d'un médecin qui pourrait lui-même poser un diagnostic avant tout, en quelque sorte, tu vois ? Donc, je sais. C'est un avis comme un autre [rire]
Interviewer	OK.
Participant 23	c'est tout ?
Interviewer	Non. Et comment est-ce que tu décrirais ton expérience de Babylon Health ?
Participant 23	c'était, je trouve, très facile d'utilisation.
Interviewer	Hm-hm.

Participant 23 c'est assez vite compréhensible, on voit très vite ou 4a veut en venir.

Interviewer Hm-hm.

Participant 23 [pause] Je pense que ça serait peut-être bien aussi, alors je ne sais pas si c'est une application nouvelle ou pas

Interviewer Hm-hm ?

Participant 23 Mais de proposer peut-être [pause] plus ? Parce que je n'ai pas regardé, à chaque fois, il y avait peut-être deux ou trois possibilités. Il y a eu une ou deux pages où il y avait quand même plus et là, pour moi, ça me parait déjà plus concret, en quelque sortes, parce que, en fait, on peut avoir plein de choses. Et surtout là, avec le temps de covid, etc., il peut y avoir, en quelques sortes, d'autres choses qui pourraient amener sur un tout autre diagnostic, tu vois, Je ne saurais pas dire combien c'était complet, parce que tu vois, mes compétences médicales elles sont assez [rire] poussées, mais ouais, par contre, disons que très facile à utiliser, Je pense que ça peut être assez utile, aussi, mais dans une certaine mesure. Il faut que ça soit très encadré et que la qualité soit bonne. Et ça, je ne suis pas capable de le juger.

Interviewer OK. Qu'est-ce que tu as le plus apprécié dans ce produit ?

Participant 23 Je pense la facilité d'utilisation. C'est très facile à utiliser. Et puis...

Interviewer Et qu'est-ce que t'as moins aimé ?

Participant 23 [pause] Qu'est-ce que j'ai le moins aimé ? [Pause] Je ne crois pas vraiment que j'ai de choses que j'ai moins aimé.

Interviewer OK. Est-ce que tu avais déjà utilisé une application similaire auparavant ?

Participant 23 Non.

Interviewer d'accord. Je vais te laisser continuer le *survey* tranquillement. Tu peux arrêter le partage d'écran.

Participant 23 Ca roule !

Participant 24

Niveau d'anglais annoncé :

Maitrise de la langue : Excellent

Trajet dans l'application : Parfait

Commentaire : le participant avait eu les mêmes symptômes le mois précédant l'entretien. Babylon a diagnostiqué comme son docteur.

Transcription de l'entretien

Interviewer	Hm-hm. Est-ce que vous pouvez revenir en arrière ? J'aimerais vous poser quelques questions.
Participant 24	OK.
Interviewer	Pourquoi est-ce que vous avez hiérarchisé ces qualités de cette manière ? Pourquoi est-ce qu'elles sont importantes pour vous ?
Participant 24	Euh, pour moi il faut que ça donne un pseudo diagnostic qui soit le plus plausible possible. Déjà que ça soit en concordance avec les symptômes et que ça soit, enfin, que ça puisse donner une réponse qui soit le plus proche possible que poserait un médecin. Pour que ça soit, pour qu'on considère ça assez fiable quoi. Et puis, bon, j'avoue que je m'en fiche un peu de la qualité esthétique de l'outil en question. Je juge plus sur l'utilité et [racle la gorge] et le fait que ça soit compréhensible aussi pour n'importe qui, quoi.
Interviewer	OK. Et comment vous décririez votre expérience de Babylon Health ?
Participant 24	Euh, bah très bien, parce que [rire] vraiment c'est une coïncidence mais j'ai eu ça le mois dernier
Interviewer	Ah !
Participant 24	Et c'était un blocage de la trompe d'Eustache. Et donc c'était pareil, j'avais mal dans la mâchoire et j'avais une oreille qui était un peu bloquée, j'avais mal à l'oreille. Donc euh je trouve que c'est bien [rire]
Interviewer	[rire] Qu'est-ce que vous avez le plus apprécié dans ce produit ?
Participant 24	Bah j'aime bien c'est que ce n'est pas - ce n'est pas comme Doctissimo où [racle la gorge] où on ne met pas un symptôme et ça nous donne tout ce que ça peut être possible, du pire au... Du coup, je trouve que, ça rend un peu, c'est quoi le mot... enfin, ça fait un peu peur, quoi. Alors, que là, ça cible beaucoup plus, en prenant en compte tous les symptômes et ça cible vachement plus ce que ça peut être. Il n'y a que deux possibilités de ce que ça peut être, deux choses assez proches. Et je trouve que c'est bien.

Interviewer	Et qu'est-ce que vous avez le moins apprécié dans le produit ?
Participant 24	Oh, bah, rien. Je ne sais pas. C'est quand même assez lent à faire, il y a beaucoup de questions. Mais bon, c'est ça qui permet d'être plus fiable. Donc c'est un point négatif, sans en être un.
Interviewer	Et est-ce que vous avez utilisé une application similaire auparavant ?
Participant 24	Non. Enfin, Doctissimo quoi, ou des trucs comme ça, mais ce n'est pas.
Interviewer	OK ouais.
Participant 24	Accessible.
Interviewer	Super. Merci beaucoup. Je vais vous laisser continuer le *survey* et vous pouvez arrêter le partage.
Participant 24	OK.

Participant 25

Niveau d'anglais annoncé : C2

Maitrise de la langue : excellent

Trajet dans l'application : Bon

Commentaire : /

Transcription de l'entretien

Interviewer	Est-ce que vous pouvez revenir en arrière, s'il vous plait ?
Participant 25	Ah, oui.
Interviewer	J'aimerais vous poser quelques questions
Participant 25	Ouais.
Interviewer	Pourquoi est-ce que vous avez hiérarchisé ces qualités de cette manière ? Pourquoi est-ce qu'elles sont importantes pour vous ?
Participant 25	D'abord parce que si ce n'est pas utile, je n'utiliserais pas l'application. Ensuite, si ça ne donne pas du contenu qui m'a l'air plus ou moins correct, ça n'a pas grand sens d'utiliser une application. Aussi, il faut que ça soit, du coup, efficace et qu'on puisse compter dessus et que ça ne va pas, tout le temps, enfin, que ça bug pas, quoi. Ensuite, « *Dependability* » je pense que c'est la même chose, enfin, que c'est une application sur laquelle on peut compter. Et que si c'est trop compliqué à comprendre, ça me donnerait moins envie de l'utiliser. Mais c'est clairement moins important que les autres choses, pour moi. La qualité du contenu, peut-être que je le mettrais plus haut, au-dessus de « *Perspicuity* », parce que c'est aussi important. Ça rentre aussi dans la confiance du contenu ou pas, et que, du coup, si ce n'est pas du contenu, enfin, très complet par exemple, ce n'est pas hyper intéressant. Après, aussi, si ce n'est pas hyper complet mais qu'il y a moyen de chercher autre part des informations. Le fait que c'est une application, enfin, attirante, je vois ça en termes de visuellement, et ce n'est pas quelque chose de primordial, à priori. Et aussi, un peu l'originalité de l'application, ce n'est pas non plus quelque chose de primordial, pour moi. Enfin, je ne sais pas si j'ai bien compris tous les mots, mais je pense ça.
Interviewer	Oui, oui.
Participant 25	OK.
Interviewer	Et -
Participant 25	Euh
Interviewer	Dites-moi.
Participant 25	Et j'allais dire que je ne sais pas si tu as d'autres questions.
Interviewer	Oui !

Participant 25 Par rapport -

Interviewer Comment est-ce que vous décririez votre expérience de Babylon Health ?

Participant 25 C'était un peu lent, mais que globalement c'était facile à comprendre et que c'était assez clair à utiliser. Mais, oui.

Interviewer Et qu'est-ce que vous avez le plus apprécié dans ce produit ?

Participant 25 Je pense, le fait que, enfin, que ça posait plein de questions et que ça allait en profondeur et que ça paraissait quand même précis, les questions que ça posait, et pas juste, enfin, « est-ce que vous avez mal à la gorge », mais plus « quel type de maux de gorge », quelque chose comme ça. Et je trouvais ça bien et, oui.

Interviewer Qu'est-ce que vous avez le moins apprécié ?

Participant 25 Le fait que ça prenne du temps, je pense. Enfin, globalement, si tu es en train d'essayer de comprendre tes symptômes, j'utiliserais plus Google et ça va beaucoup plus vite, même si ce n'est pas. Enfin, même si le temps, je prends le temps de vérifier si les sources m'ont l'air fiables ou pas. Et du coup, peut-être que ça pourrait s'équilibrer d'une certaine manière, de prendre plus de temps de répondre à des questions, mais d'avoir des sources des sources un peu plus fiables.

Interviewer D'accord. Et est-ce que vous aviez déjà utilisé une application similaire auparavant ?

Participant 25 Ah, non.

Interviewer D'accord. Je vais vous laisser terminer le *survey* tranquillement. Vous pouvez arrêter le partage.

Participant 25 OK.

Participant 27

Niveau d'anglais annoncé : B2

Maitrise de la langue : excellent

Trajet dans l'application : Bon

Commentaire : /

Transcription de l'entretien

Interviewer	Est ce que tu peux revenir en arrière ? J'aimerais te poser quelques questions.
Participant 27	Yes.
Interviewer	Pourquoi est-ce que tu les as hiérarchisés de cette manière ? Pourquoi est-ce que ces qualités sont importantes pour toi ?
Participant 27	**P**our moi, j'ai mis en premier « *Dependability* » selon la définition qui est marquée ici « *the quality of being trustworthy and reliable* » parce que pour moi je dirais que si c'est une application liée à la santé, le plus important c'est qu'elles soient « reliable » et « *Trustworthy* », plutôt que par exemple, j'ai mis « *Novelty* » en dernier parce que, en soi, je ne pense pas que les gens qui iront sur ce site, je ne pense pas qu'ils cherchent beaucoup de nouveautés en général. J'imagine qu'ils auront des symptômes assez basiques et que s'il avait vraiment des symptômes très, très importants ils iraient peut-être directement aux urgences. *I hope so*. Après, j'ai mis « *Perspicuity »* [rire] parce que, aussi, c'est important de pouvoir avoir un diagnostic mais il faut qu'il soit vraiment clair et facile à comprendre, parce que sinon le site n'a pas beaucoup de... intérêt, je trouve pour le patient. Et puis, j'ai mis « *Trustworthiness of Content »* qui, pour moi, en soi, se rapporte un peu à « *Dependability* ». Donc, c'est encore l'idée de pouvoir genre, en soi, avoir confiance dans le diagnostic qu'on te donne. Puis « *Attractiveness* », je l'ai mis en quatrième, parce que je considère que si un site n'est pas très « *attractive* », il y a peu de chances qu'on utilise. Et, moi j'avoue que j'aimais bien le *layout* du site et puis j'aimais bien les couleurs choisies. C'était mauve. Après, je ne sais pas si les mecs aimeraient bien, mais bon voilà. Puis « *Usefulness* » moi j'avoue, je sais... En fait, honnêtement, j'ai mis ça en cinquième parce que je ne sais pas si j'utiliserais beaucoup ce genre de sites, parce que... J'ai plus tendance soit à regarder vite fait regarder sur internet ou alors directement prendre rendez-vous avec mon médecin généraliste. Je trouve plus simple. Parce que, au final, après avoir fait, après avoir répondu au questionnaire, je vois quand même qu'il propose

quand même de prendre rendez-vous avec ton GP. Donc je me dis, au final, « est-ce que tu ne pourrais pas juste prendre directement un rendez-vous avec ton GP ? » ça serait plus facile, ça enlèverait une étape. Et puis « *Efficiency* » et « *Content Quality* », je ne savais pas trop où les mettre, je les ai un peu mis tous les deux, genre, l'un au-dessus de l'autre sans vraiment beaucoup de raisons. Au final, « *Content Quality* », *yes*, mais, quand j'ai vu les résultats c'était juste une toute, toute petite description. C'était juste le nom de la possible maladie ou la cause de symptômes, juste une petite discussion ce que c'était. Donc au final, je ne sais pas si le « *content* » était très qualitatif. Donc, voilà.

Interviewer OK.

Participant 27 Je ne sais pas si tu as d'autres questions [rire].

Interviewer Oui, j'ai quelques-unes encore. Comment est-ce que tu décrirais ton expérience de Babylon Health ?

Participant 27 Assez bonne. Je dirais que les questions étaient assez pertinentes. Il n'y en avait pas trop peu, ni trop... trop en général. Parce que, j'avais un peu peur qu'il y en ait beaucoup trop et qu'il y ait des questions qui ne soient pas trop liées, en fait, au symptôme principal qu'il y avait. Donc, la douleur dans les oreilles. Finalement, tout était un peu lié, je trouve, à la tête. Donc, il y avait, genre, les maux de tête, les muscles et tout donc c'était... C'est bien. Et aussi, j'avais... peur qu'il n'y en ait pas assez et qu'il te donne directement la réponse la plus facile, du genre « ah bah t'as mal aux oreilles, ben va juste voir ton médecin, ce n'est pas grave, on verra ». Enfin, tu vois quoi. Il y avait quand même une proposition, j'imagine que c'est pas mal. Les questions, je trouve, étaient assez simples à comprendre. Après, bon, moi je devais réfléchir, parce qu'il y avait le petit, euh, la petite description que je devais aller relire de temps en temps. Mais, je que si tu devais le faire pour toi même, en connaissant tes symptômes, c'est assez simple. Donc c'était une expérience assez plaisante, je trouve. Et voilà, moi c'est vrai que mon seul problème, je dirais, que c'est plus l'utilité, tu vois, de l'application. Mais, dans le *user experience*, je trouve qu'il est très bien.

Interviewer OK. Et qu'est-ce que tu as le plus apprécié dans ce produit ?

Participant 27 Pour moi, je dirais, c'est sa simplicité. Quand il demandait de cliquer sur certains symptômes, ou, euh, quoi, il n'y avait pas des longues définitions, des longs paragraphes sur ce que tu pouvais avoir. C'était assez straight to the point, genre, oui/non, « *I don't know* » ou alors « j'ai mal là, mal là, mal là », « *none of the above* ». Donc je trouve que c'est pas mal. Après, c'est

	vrai que je me demande si les gens appuyaient sur « *I don't know* », est-ce que ça changerait énormément les résultats ? Parce qu'alors, évidemment, ça donne beaucoup moins d'informations à la plateforme pour te donner un diagnostic. Donc voilà.
Interviewer	Et qu'est-ce que t'as le moins apprécié dans ce produit ?
Participant 27	Ce que j'ai le moins apprécié dans ce produit, je dirais c'était [pause] Peut-être dans les résultats, la page de résultats, le fait qu'il n'y ait pas assez d'informations, justement. Je pense que j'aurais voulu avoir plus d'informations sur les deux causes principales qu'ils mettaient en avant. Quitte à avoir, au final, peut-être un lien qui renvoie à un autre site, ou, enfin, à une page de leur propre site. Mais je pense que ça aurait été pas mal. Ou un peu plus vérifier si ça peut être l'un ou l'autre. Et je me dis que les personnes qui vont utiliser ce site sont quand même assez intéressées par rapport à ça.
Interviewer	Hm-hm
Participant 27	Donc voilà.
Interviewer	Est-ce que tu as déjà utilisé une application similaire auparavant ?
Participant 27	**N**on. Jamais.
Interviewer	D'accord. Top. Je vais te laisser continuer et terminer le questionnaire tranquillement. Tu peux arrêter le partage.

Participant 28

Niveau d'anglais annoncé : C1

Maitrise de la langue : excellent

Trajet dans l'application : good

Commentaire : elle m'a demandé s'il fallait faire un covid check et a décidé de le faire.

Transcription de l'entretien

Interviewer	Est-ce que tu peux revenir en arrière ? J'aimerais te poser quelques questions. Ah, t'as pas ton micro par contre.
Participant 28	Désolée, voilà.
Interviewer	Donc, je voulais savoir, pourquoi est-ce que tu les as hiérarchisés de cette manière ? Pourquoi est-ce que ces qualités sont importantes pour toi ?
Participant 28	Je pense que c'est... Attends. Pourquoi est-ce que j'ai mis que le produit était attractif en premier, tu veux dire ?
Interviewer	Ouais, pourquoi est-ce que tu trouves que l'attractivité c'est la chose la chose la plus importante dans une Health app ?
Participant 28	Ah, pardon, j'avais mal compris, en fait. Je pensais qu'il fallait classer en fonction de ce qu'on pensait que, quel était le trait le plus important dans l'application, et pas pour nous, en fait. Donc
Interviewer	D'accord ouais.
Participant 28	C'est moi qui -
Interviewer	Pas de soucis.
Participant 28	[remet dans l'ordre] Donc, maintenant pour répondre à ta question, moi j'aurais tendance à classer les qualités de l'application dans cet ordre-là, parce qu'avant toute chose, il faut que je me sente à l'aise avec les informations qui me sont données. Donc il faut que je puisse avoir confiance au diagnostic que je reçois, donc il faut que ça soit, enfin, des informations sûres et de confiance. Et puis, en troisième position, le fait que ça soit clair, ouais, que ça soit des informations claires et concises, c'est important aussi. Bon, vu que je ne suis pas un professionnel médical, il faut qu'en quelques mots, ou en tout cas, en peu de phrases, je puisse comprendre ce qui m'arrive ne tout cas, avant même que l'application soit révolutionnaire ou quoi que ce soit. Si je ne comprends pas, en fait, ça ne sert à rien pour moi que ça soit quelque chose de totalement nouveau, ou voilà. Donc voilà, je pense que la confiance, ou la sureté des informations et l'efficacité aussi. Parce que, bon, mine de rien, si j'ai une information mais que ça ne m'aide pas à savoir quelles

sont les prochaines étapes que je dois prendre, ça ne servira pas beaucoup. Donc voilà. Et puis, pour ce qui est de l'attractivité, on parle de soins de santé, ou en tout cas une consultation, le fait que ça soit sexy ou quoi, c'est le dernier de mes problèmes, en fait. Bon, voilà. C'est pour ça que je classerais ça comme ça.

Interviewer OK. Et comment est-ce que tu décrirais ton expérience de Babylon Health ?

Participant 28 J'aurais tendance à dire que c'était un peu froid. Bon il n'y a pas ce contact humain qui est nécessaire pour vraiment pouvoir expliquer de manière claire, au-delà des symptômes, l'état d'esprit dans lequel on se trouve, pour faire un diagnostic, je trouve. Et aussi, des fois, il fallait quand même que je lise bien chacune des options pour comprendre si ça s'applique à mon cas ou pas et bon, j'avais un peu des difficultés à comprendre exactement si je devais mettre oui ou non pour l'une des options. Ouais, sinon, c'était relativement rapide, donc ça, c'est quelque chose de bien. Et clair aussi. Clair. Donc, au niveau de la plateforme, au niveau visuel, en tout cas, je trouvais que ça donnait bien. C'était assez épuré et il n'y avait pas des informations à droite à gauche. Enfin, le site était assez bien agencé, je dirais.

Interviewer OK. Est-ce que le produit correspondait à tes attentes ?

Participant 28 Pas vraiment. En tout cas, pas pour le résultat que j'ai reçu. Il y avait pas mal d'options, différentes options de maladies que je pouvais avoir et chacune me référait à un médecin. Donc au final, je me dis, si c'est pour toujours me référer toujours à un médecin, à quoi 4a sert de l'avoir fait ? Autant, j'aurais peut-être mieux fait de directement appeler pour consulter, en fait. Donc, de ce point de vue-là, je pense que c'était un peu en dessous de mes attentes.

Interviewer OK. Qu'est-ce que tu as le plus apprécié dans ce produit ?

Participant 28 L'idée, en fait. Le concept de pouvoir soi-même déterminer, un peu, qu'est-ce qu'on pourrait avoir. Enfin, le concept de pouvoir se dire, « OK, je ne me sens pas bien, peut être que à la fin, je vais savoir exactement ce que j'ai », plutôt que de devoir me déplacer dans les transports en commun pour voir un médecin. Donc, en fait, je trouve que l'idée est quand même assez intéressante. Au-delà du résultat, bien sûr.

Interviewer Et ce que t'as le moins apprécié ?

Participant 28 Bonne question. Le temps que ça a mis, pour être honnête [rire]. Bon, c'est un peu contradictoire, mais je m'explique. En général, quand on parle face à face avec un médecin, je trouve que c'est plus facile de s'ordonner, de s'organiser. Mais là j'avais du mal, je devais à chaque fois revenir sur l'autre plateforme pour

	chercher mes symptômes, puis revenir pour indiquer. Donc, bon, il y avait beaucoup de va et vient, en fait. Et certaines questions je trouvais, se ressemblaient quand même un peu. Donc, je me demandais « si j'ai déjà répondu à ça, pourquoi c'est reposé sous une autre forme ? », donc, peut-être, ouai voilà, le fait que ça ait mis un peu plus de temps que ce que j'aurais cru.
Interviewer	D'accord. Et, enfin, est-ce que tu as déjà utilisé une application similaire auparavant ?
Participant 28	Non. Je ne pense pas. En tout cas, pas pour les soins de santé, mais ce qui est plutôt pour les assurances, ou autres. Il y a un peu des plateformes qui te redirigent pour savoir à quel service tu dois t'adresser exactement. Mais pas pour avoir un diagnostic précis à la fin. Donc j'aurais tendance à dire que non, je n'ai pas vraiment fait l'expérience d'une plateforme similaire.
Interviewer	OK. Super. Je vais te laisser continuer et terminer le questionnaire tranquillement.
Participant 28	OK.
Interviewer	Tu peux arrêter le partage.
Participant 28	D'accord.

Participant 29

Niveau d'anglais annoncé : C2

Maitrise de la langue : excellent

Trajet dans l'application : excellent

Commentaire : Le participant 29 était très enthousiaste et taquin

Transcription de l'entretien

Interviewer	Est-ce que tu pourrais revenir en arrière ? J'aimerais te poser quelques questions.
Participant 29	[rire] Oui.
Interviewer	Pourquoi est-ce que tu les as hiérarchisés de cette manière ? Pourquoi est-ce ces qualités sont importantes pour toi ?
Participant 29	Hm, OK. Bon, je peux te dire pourquoi « *Attractiveness* » je l'ai mis en dernier, parce que je pense que pour une application qui concerne la santé, ce n'est pas forcément le design qui va me convaincre, mais plutôt la qualité de l'information et à quel point c'est, comment on dit [pause] pertinent ou non et si tu peux avoir confiance en ce qu'on te dit ou pas. Sachant que quand même, je partirais, enfin l'utilisation pour moi de cette application ça ne serait pas pour, vraiment pas pour remplacer un médecin mais plutôt [pause]. En fait, même. Ouais. Comme tu utiliserais Google, par exemple, ou d'autres sites comme Doctissimo, ça, par exemple, ce n'est pas possible [rire] Donc si, ça c'est une alternative qui fait du sens, parce que le diagnostic qu'on te propose est réaliste, pourquoi pas. Dou le fait que je place « *Trustworthiness of Content* » en premier et « *Attractiveness* » en dernier. Mais, ensuite, en vrai, j'étais obligée de les classer les huit ? Non. Oui ?
Interviewer	Oui.
Participant 29	Oui, donc voilà. [Pause]
Interviewer	OK.
Participant 29	Est-ce que ça t'aide ma réponse ou pas ?
Interviewer	Oui, oui, je te laissais un peu de temps parce que tu m'avais l'air pensive.
Participant 29	[rire] OK
Interviewer	Et sinon, comment est-ce que tu décrirais ton expérience de Babylon Health ?
Participant 29	Alors [pause]. J'ai vu, enfin, à quel moment, enfin, je voulais arriver à ce « *jaw pain* », tu vois ? [rire] Mais, du coup, quand même [pause] Vu qu'il y avait un scénario, j'ai essayé quand même de rester au plus proche de ça, et je me dis, dans le cas où tu n'as pas de scénario, tu peux prendre en compte plein de choses qui se passent dans ton corps à ce moment-là et du coup, comment tu fais le tri entre ce que sont vraiment des

	symptômes de ce que tu peux avoir, et des choses biologiques qui se passent dans ton corps ? [rire] c'est une réflexion que j'avais. Mais, sinon, j'ai trouvé que c'était extrêmement facile à utiliser, que les questions à chaque fois étaient claires et que les propositions elles étaient assez, elles aiguillaient vers différents types de douleurs, par exemple. Et du coup c'était assez facile de choisir la bonne douleur. [rire] Est-ce que ça fait du sens ce que je dis ? Je ne sais pas, mais voilà. Après, j'ai trouvé qu'il y a eu quand même un certain nombre de questions, ce qui bon, prouve aussi, enfin, ce qui en tout cas donne de la confiance dans le diagnostic. Et, après, c'était quand même relativement rapidement fait et les diagnostics, j'ai trouvé aussi ça intéressant la façon dont c'était fait, avec les risques, etc. Et, malheureusement pour moi, je dois aller chez le médecin rapidement [rire]. Mais, voilà. Et aussi, c'est indiqué quand même que ça ne remplace pas u médecin, ce qui est bien aussi. Mais voilà. C'est très *user friendly*, en tout cas. Et facile à utiliser et compréhensible, voilà.
Interviewer	OK. Est-ce que ça correspondait à tes attentes ?
Participant 29	Ouais [rire] En fait, quand j'ai lu la liste de mes symptômes, je me suis dit « ah, serait-ce une otite ? » [rire]. Et du coup, je pensais en tout cas, on allait partir sur quelque chose comme ça, mais je n'aurais pas pensé à la sinusite, tout de suite. Mais il y avait l'histoire de la *jaw*, ce qui est, voilà.
Interviewer	Qu'est-ce que tu as le plus apprécié dans ce produit ?
Participant 29	Je pense vraiment la facilité d'utilisation, parce que je m'attendais ce qu'il y ait, forcément, des questions qui me soient posées par rapport à mes symptômes, mais pas que ça soit aussi, en tout cas que ça paraisse aussi bien fait, en fait. La ... Ouais.
Interviewer	Et qu'est-ce que tu as le moins aimé ?
Participant 29	Ouh ! Je n'ai pas vraiment de critique à faire, en fait. [pause]
Interviewer	OK. Est-ce que t'avais déjà utilisé une application similaire auparavant ?
Participant 29	Non, jamais. First time.
Interviewer	Super. Je vais te laisser continuer et terminer le questionnaire tranquillement. Tu peux arrêter le partage.
Participant 29	OK.