How Girl Effect built a chatbot

By Calum Handforth and Kecia Bertermann
Girl Effect is a creative non-profit that uses media and mobile technology to empower girls to change their lives. Set up by the Nike Foundation in 2004, Girl Effect is active in 66 countries and has reached and engaged with more than 48 million people through youth brands and mobile platforms. Our work helps girls to express themselves, value themselves, and build the relationships they need. Girl Effect is a UK Registered Charity (1141155).

This paper would not have been possible without the contributions of Isabelle Amazon-Brown, Ximena Benavente, Cathy Bussey, Adam Cadman, Sara El Dabi, Lindsey Evans, Amy Green, Patricia Hutny, Khwezi Magwaza, Jonathan McKay, Leah Muwanga-Magoye, Linda Raftree, Lorya Roblin, Neville Tietz and Kira Vosper - and with a particular thank you to Jessica Heinzelman for her guidance throughout the project.
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Key takeaways</strong></td>
<td>3</td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
<td>4</td>
</tr>
<tr>
<td>About the chatbot</td>
<td></td>
</tr>
<tr>
<td>The development and testing process</td>
<td></td>
</tr>
<tr>
<td><strong>Insight one: Build for the user</strong></td>
<td>10</td>
</tr>
<tr>
<td>Understanding users</td>
<td></td>
</tr>
<tr>
<td>Delight users</td>
<td></td>
</tr>
<tr>
<td>Privacy and security are non-negotiable</td>
<td></td>
</tr>
<tr>
<td><strong>Insight two: Chatbot writing is unique</strong></td>
<td>14</td>
</tr>
<tr>
<td>Understanding the value of content</td>
<td></td>
</tr>
<tr>
<td>Build a content process</td>
<td></td>
</tr>
<tr>
<td>Humans are still important!</td>
<td></td>
</tr>
<tr>
<td><strong>Insight three: Using what works in the tech world</strong></td>
<td>18</td>
</tr>
<tr>
<td>Using Sprints to build a Minimum Viable Product</td>
<td></td>
</tr>
<tr>
<td>Rethinking success</td>
<td></td>
</tr>
<tr>
<td>Test, learn, and iterate</td>
<td></td>
</tr>
<tr>
<td><strong>Insight four: Changing how we measure</strong></td>
<td>25</td>
</tr>
<tr>
<td>Making measurement strategic</td>
<td></td>
</tr>
<tr>
<td>The realities of measuring in a digital environment</td>
<td></td>
</tr>
<tr>
<td>Thinking broader with measurement</td>
<td></td>
</tr>
<tr>
<td><strong>Insight five: Technology can be complex</strong></td>
<td>29</td>
</tr>
<tr>
<td>It’s about more than the technology</td>
<td></td>
</tr>
<tr>
<td>Don’t forget about offline</td>
<td></td>
</tr>
<tr>
<td>Launch is not the end!</td>
<td></td>
</tr>
<tr>
<td><strong>Reflections and Conclusion</strong></td>
<td>33</td>
</tr>
</tbody>
</table>
Key takeaways

Workshops with adolescent girls in South Africa and the Philippines highlighted a need for accurate information, without judgement. We therefore developed a chatbot to provide girls with a private space in which to ask sensitive and important sexual and reproductive health questions.

Although Girl Effect is a digitally-savvy organisation that reaches millions of girls around the world through social media, mobile platforms, and youth brands, we learned a lot from developing the chatbot, and are keen to share five insights with others exploring the potential of this technology.

1. Insight one: Build for the user
The user experience is sacrosanct. All decisions need to be made with users at the centre. This begins with understanding user needs and realities, including whether or not a chatbot is a channel with which they can engage. We had to focus on developing engaging and personable content to delight users, to make sure they found the information that they were looking for, and to remain engaged with the chatbot. All of our efforts were driven by respecting users’ privacy and security: non-negotiable aspects at Girl Effect.

2. Insight two: Chatbot writing is unique
We found that chatbot content is very different to other forms of online, and offline, content. Building a content development and management process was essential, allowing us to manage content from creation to uploading, and refinement. Reflecting the always-on nature of the chatbot, we had to write content that both provided users with immediate answers, and encouraged them to explore the chatbot. However, reflecting on the importance of the user experience, content needs to be streamlined: any superfluous steps risk losing users. Human input is also important for some users: for moderation, for safeguarding, and for signposting.

3. Insight three: Using what works in the tech world
We drew on best practice from the technology sector in developing the chatbot. In particular, we learned the value of using time-defined Sprints to focus our efforts, and the importance of building a Minimum Viable Product. This meant being comfortable with testing, learning, and iterating. We had to recognise that we would never have a perfectly finished product, and refinement in collaboration with users was essential. Similarly, we had to rethink our traditional definitions of ‘success’ - learning from other technology interventions, and aligning these approaches with international development best practice.

4. Insight four: Changing how we measure
Chatbot users need answers to their questions, now. They demand immediate value, or they’ll go elsewhere. Reflecting this, we had to engage with the realities of measuring in this setting. Extraneous, transactional, or abrupt measurement had a real and negative impact on the user experience. We had to think broadly with measurement: look to back-end metrics, dashboards and other analytics tools, consider digital proxies, but also test the value of more ‘traditional’ measurement approaches. We had to take a strategic approach to measurement and recognise what had to be measured immediately, identify measurement priorities for the future, and drive actionable decisions that improved the user experience and value of the chatbot.

5. Insight five: Technology can be complex
We noticed it was difficult not be led by and base all our decisions on the technology. Instead, the chatbot was a tool to reach our users and had to be driven by their needs. We continuously reflected to ensure this remained the case. Similarly, we couldn’t just rely on technology. For the sensitive and important topics that users were exploring, we had to have human intervention. This included moderation, but also founding the entire project on extensive in-person user engagement at the outset. Offline still has a lot to offer here, particularly in exploring the needs and realities of users. Finally, we also quickly learned that launch is not the end! Bug-fixing, refinement, and bigger changes based on real and varied user journeys are significant and longer-term undertakings.

This is a very brief summary of all that we learned, and we hope that the following pages provide further guidance and insight to international development organisations who are exploring the potential of chatbots. From the outset, we had very little guidance or best practice to follow. Most resources were very focused on the role of chatbots in the private sector, and few offered real and practical advice. We hope that this paper can play a role in filling this gap.
Introduction
Girl Effect uses media and mobile technology to empower girls to change their lives. As part of this approach, we’ve been exploring the potential of chatbots to engage girls where they are: on social media, and on their mobile phones.
Through our years of research with girls all over the world, they have consistently highlighted a need for a private and non-judgemental environment in which to ask sensitive questions. We thought that a chatbot, as an online and private channel, may be a potential solution to empowering girls by providing these answers. Building on this consideration, we conducted in-person and chatbot-focused research with groups of adolescent girls in South Africa and the Philippines.

These girls highlighted that they wanted confidential answers to a range of sensitive sexual and reproductive health questions. They noted that they often couldn’t discuss these topics offline, particularly due to social stigma and a lack of trust in others. In addition, when they did engage offline, they were often given unhelpful or incorrect information. We also saw that mobile phones were a central tool in their lives - and often a key source of information. This confirmed the value of exploring the potential role of a chatbot in providing girls with the space to answer their questions.

More widely, this research – and previous Girl Effect work – identified that girls have a complicated relationship with their peers. Trust is often lacking in these relationships, a factor exacerbated by what is seen as a ‘gossiping’ culture. This complexity further prevents girls from getting the answers they need to their important questions. They want a straightforward way to get information on difficult topics, through a relatable, entertaining and non-judgmental channel.

Beyond the importance of mobile phones for this audience, we have also seen a significant increase in the usage of messaging applications – notably Facebook Messenger and WhatsApp by these users (as well as across the international development sector1, including to communicate with donors and supporters3). We wanted to go where our users were, provide the support that they were demanding, and deliver engaging and private information that wasn’t available on our other channels3. From the qualitative and quantitative feedback recorded during the testing phase, we demonstrated that a chatbot was a valuable channel for the audience with which we engaged (see Figure 1).

---

1. Messaging Applications for International Development A research initiative from the Digital Impact Alliance (DIAL) and Echo Mobile, available here: https://messengers.digitalimpactalliance.org
2. 2018 Global NGO Technology Report, available here: techreport.ngo
3. Girl Effect recognises the importance of quality content and social support as essential for encouraging behaviour change. Girl Effect uses this ‘content + conversations’ model in the Springster site as well as across other brands and products to ensure that girls have access to accurate, girl-friendly information and opportunities to share information and practice new skills in a supportive environment.
About the chatbot

The chatbot is currently hosted in Facebook Messenger. We selected this channel due to the comparatively lower barriers of this platform in developing a working prototype. This decision is discussed further in this paper.

Girls open Facebook Messenger to start interacting with the chatbot. The chatbot is in a narrative format - similar to a Choose your own adventure story, with the girl moving through the chatbot by clicking on options or using a numbered menu (see Figure 2). Chatbot content is split into three main areas: navigation and engagement, message sets, and quizzes. Message sets provide content to improve girls’ understanding, whilst quizzes provide an opportunity to test this (or prior) knowledge. Both message sets and quizzes focus on improving girls’ knowledge of three main topics: love and relationships, the ‘basics’ of sex, and how to have safe sex.

The chatbot was built using RapidPro – an open-source platform launched by UNICEF, which can be used to build mobile applications and services. This is the ‘back-end’ of the chatbot. Although users currently interact with the chatbot through Facebook Messenger, RapidPro is platform-agnostic and therefore the chatbot can be connected to other channels such as WhatsApp. The chatbot incorporates some basic artificial intelligence - largely word matching - but mainly comprises of quick reply buttons.

The back-end of the chatbot tracks user journeys and user responses - both to improve her future experience, and to provide the Girl Effect team with the data needed to meet the requirements of users. This setup also allows effective moderation - with danger or ‘trigger’ words flagged in real-time, allowing a team member to talk directly to the user in order to signpost her to the most appropriate support service. The intricacies of moderation are discussed in detail later in this paper.

The chatbot was initially designed to feature two different personas: Big Sis, and Springbot (a summary of both personas can be found on the next page). An aspect of the testing phase was to determine which persona was preferred by users - using data to drive decision-making. During the initial testing phase, in which more than 2,700 users interacted with the two personas, the Big Sis persona was the clear favourite. This was the case across all four countries, and therefore the Springbot persona was retired for future phases.
Chatbot personas

**Big Sis** is old enough to be on the cusp of having to do some proper ‘adulting’, but still strongly connected to youth culture. In her early twenties, she has been through the embarrassment of puberty, the feeling she’s cracked it as she turned 18, to figuring out that there’s still lots to learn from other sources, which is where she is now. She is here to make sure girls feel free to be curious, and to provide them with information that gives them the confidence to trust themselves and to act independently. She wants to be the big sister every girl wishes she had: a fun, intelligent best friend who is always there for you; keeps your secrets without judging you; and, most importantly, would let you borrow her clothes... if only she wasn’t a chatbot!

**Springbot** is an interactive guidebook brought to life. Intentionally gender-neutral, Springbot has collected knowledge and experience from girls all over the world. Springbot feels great when a girl asks something new as it is a chance to add to the wealth of information that Springbot has amassed. This new knowledge can then be useful to help other girls. Springbot is always learning, and always keen to find answers to the questions girls are asking. Springbot wants to inspire and inform girls so that they can act independently and make the best choices for themselves in relation to sexual and reproductive health. Springbot is for every girl who needs a voice of reason, information and reassurance. A voice that isn’t afraid to tell it like it is, but understands that you might be afraid to ask.
The development and testing process

Over the course of 2018, we conceptualised, developed, tested and evaluated a chatbot. The entire process is set out in Figure 3.

Building on existing research and engagement with girls, Girl Effect conducted an initial foray into the chatbot world - confirming that it was a potentially useful channel. Following this, we identified the sexual and reproductive health knowledge needs of girls through analysis of Springster website and Facebook comments - and in-country research in the Philippines and South Africa. This in-country research was also an opportunity to test, in person, the salience of a chatbot.

In order to move forward quickly, we borrowed best practice from the tech sector and ran an intensive ‘Sprint’. The benefits, and nuance, of this process are discussed extensively in this paper. The Sprint allowed us to develop a prototype chatbot - featuring the Big Sis and Springbot personas - and provide basic content and quizzes. We then tested this with real users.

We recruited users through Facebook adverts - allowing control over the number of users engaging with the chatbot, and enabling testing without large and public promotion. We ensured that each iteration of the chatbot remained live, for this audience, for a long enough time to learn what was and wasn’t working. The testing period was sufficient to see user behaviour and engagement over a suitably realistic time-frame: including weekdays, weekends, and different times of day. Chatbot testing was split into two phases.

The first phase was a five-week test of the chatbot with real users across four countries: Nigeria, the Philippines, South Africa, and Tanzania. This focused on exploring the need for a chatbot, the salience of content, and understanding the user experience - including the good and the bad. The chatbot reached 2,767 users (including returning users), mostly within Springster’s age and vulnerability targets. The majority of users were based in Nigeria and the Philippines. Girls were overwhelmingly satisfied with the chatbot.

*The target age range is 15-17. Vulnerability targets include: out of school (or behind in school); marginalised ethnic/religious group; married early and/or given birth/have been or currently is pregnant; household poverty (measured either through access to improved water source/sanitation or World Bank poverty measure of $3.10/day or less); unstable parental relationship (e.g. orphan, does not live with parents); disability.
They felt comfortable sharing, and explored all types of content, with a particular interest in content focusing on relationships. Similarly, girls navigated the chatbot based on their personal interests and needs. These varied widely. The testing allowed us to gather quantitative and qualitative data on user behavior, including preferred persona, high demand topics, and user pain points.

We then ran a second Sprint. This aimed to address priority issues observed during the initial testing - including improving user retention and engagement, tackling challenges with quick replies and navigation, and reducing our higher than usual user acquisition costs compared to other Springster channels. This was followed by a second testing phase, to put these improvements into practice. This testing phase allowed us to see the impact of these optimisations in order to strengthen the proof-of-concept. We wanted to confirm that the chatbot was a relevant channel for girls, and to demonstrate the efficacy of a data-driven and iterative approach.

We’re now moving forward and exploring how to build on the chatbot that we have developed. This includes considering how we can best use the chatbot to meet the needs of girls. We’re also keen to support other organisations in exploring the usefulness of chatbots in international development. As part of this, we’re publishing this paper, to share what we learned.
1

Insight one: Build for the user

When working in a digital setting on products such as chatbots, all decisions must be driven by the needs - and for the benefit - of users. In order to do this, we needed to understand our user, to build a positive experience for them, and to prioritise their privacy and security.
Understanding users

We needed to understand the needs, requirements, and realities of our users. This included their existing sources of information, communication preferences – such as timing, format, language, voice, and medium of communications – and wider context, such as device type, digital and functional literacy, accessibility of technology (including gender norms that may limit regular access to devices), and the wider suitability of the channel - for example, considering broader factors such as data costs.

Fundamental too, was understanding the need for a chatbot. Was it an effective channel to engage with our users, and would it allow us to achieve our intended impact? This included understanding the role of a chatbot – whether for transactional interactions, or for building a relationship with users. Chatbots were also a very new approach to many of our users, and this shaped how we developed the entire experience.

With regard to the latter point, we needed to ensure that users understood that the chatbot was not a real person. These efforts were possibly complicated by our focus on providing a persona, and associated background for our chatbot. It also affected more technical considerations including identifying the right interaction speed and volume of questions – particularly in order to avoid probing excessively on very delicate topics. We also had to find the correct tone of voice, and ensure that automated responses – when using low-complexity Artificial Intelligence (AI) – were suitably empathetic.

We discovered that this can be a very broad area - including learning that the meaning of emojis (and rating systems, for feedback) can vary considerably across countries. As we expected, many girls were also not at all familiar with chatbots. This inexperience resulted in positive interactions - for example, users wishing the chatbot a ‘goodnight’ or expressing familiarity (see Figure 4, below) - but also more angry responses. Several users did not engage well with the pace of messages, the always-on aspect of the channel, or repeated prompts. We learned the importance of incorporating a pause or stop option more consistently, offering a deeper explanation of what a chatbot is for those users not engaging with the ‘Quick Reply’ setup, and also creating a more diverse library of responses to functionality questions.

Overall, we had to maintain a constant focus on user needs and realities – sometimes at the expense of meeting our needs and priorities associated with this project. Sometimes the two aligned, but we learned to never prioritise our needs at the expense of our users.

Figure 4. 90% of girls feel comfortable sharing with the chatbot and some express familiarity by calling it by name

“What can I do to look more attractive big sis?”

“But big sis hw will I know if he truly love me too”

“Hey Springbot ... is it true that you can’t break virginity with a condom on?”

“Hi Springbot ... how do you tell your boyfriend you don’t want to have sex with him?”

243 Direct Mentions of ‘Big Sis’ from 171 Users

29 Direct Mentions of ‘Springbot’ from 27 Users

*Girl Effect has undertaken significant research into girls’ access to technology. This includes a 2018 study in partnership with the Vodafone Foundation: A global study of girls’ access to and usage of mobile, told through 3,000 voices available here: https://www.girlsandmobile.org
Delight users

User experience is paramount. All of our decisions had to be made with our users at the centre. This included making sure content was accessible and conversational – and that we were not broadcasting to users as this wasn’t the purpose of our chatbot. We learned that chatbot users can be incredibly transient – often only engaging when requiring help, or assistance; and demanding a solution or answer with minimal interaction.

We found that although 59% of our users interacted with the chatbot for less than one hour, their completion rates across content were high (see Figure 5, below). Further analysis of user engagement highlighted that they engaged with both message sets and quizzes (Figure 6). This highlighted the importance of making things as simple as possible, and minimising the number of steps that a user had to take in order to gain value of some type.

Every additional step in the user pathway offered the potential for someone to get confused, bored, or frustrated – and to stop engaging. This was a primary reason why we elected to not make the persona selectable in the chatbot, as the user was being asked to take several further steps before gaining any direct value – in particular, not receiving any answers to her questions.

Quality Assurance (QA) was also essential to deliver an experience that our users deserved. We QA’ed the chatbot extensively before piloting and launch; and then continued refining as new issues were identified. This included QA’ing within Girl Effect, and with our technology partners. Unfortunately, we were not able to test with end-users, but would recommend doing this if at all possible. We were also unable to test the many different hardware and software configurations used by girls. Therefore, during the post-launch testing period - when ‘real’ users interacted with the chatbot - we identified some key functionality issues that significantly damaged the user experience.

In particular, 27% of all users could not see Quick Reply buttons due to the application version or handset that they were using. Other users also weren’t seeing any buttons, and had to refresh their inboxes to receive ‘automated’ replies. These issues appeared to be most common amongst users with the Facebook Lite application, those using third-party applications, and those using older handsets. In response, we created an alternative to Quick Replies using numeric menus ahead of the second phase of testing.

Although this process was frustrating, for both us and our users, all of this was very valuable learning. These issues significantly affected the user experience, but we were able to identify and fix them ahead of a future larger-scale launch.

Figure 5. Completion rate by message set. Completion of content was high across all markets and message sets (N.B. lower amount of users in South Africa)

Figure 6. Neither message sets nor quizzes had consistently higher drop-off or completion rates
Privacy and security are non-negotiable

The chatbot was built and tested in summer 2018, during implementation of the European General Data Protection Regulation (GDPR). Although the wider discussion in the sector regarding GDPR was helpful in guiding some decisions, data privacy and security are already core values at Girl Effect*. These were similarly essential priorities throughout the development of the chatbot. This included consideration of the most suitable platform, modifying our Privacy Policy to include aspects of the chatbot that are different to our other channels, and interrogating the privacy policies of third party organisations, such as analytics providers.

We know that privacy has particular relevance to building trust – notably with regard to safeguarding girls who are raising very difficult topics. Chatbot development also prompted an initial tension between data privacy and user-centred design. Building a chatbot required us to understand trends at the largest scale – leveraging Big Data and other insights in order to see what was and wasn’t working. However, effective iteration of the chatbot was also often driven by the experiences and frustrations of individual users. We had a small group of appropriately vetted team members able to access individual conversations, through Facebook Messenger, to drive progress to benefit thousands of users and to respond to safeguarding issues - discussed in more detail, later.

Our privacy and security priorities also occasionally actively conflicted with our opportunities to learn. For example, users being able, and understanding how, to delete their conversations and associated data damaged our data collection efforts and, in future, may skew our insights if the number of such users is sizeable. However, to safeguard users, particularly vulnerable users, and those who may be using shared devices we saw it as essential that they were not put at risk in seeking information and answers to their important questions.

Overall, having trust as the foundation to user engagement was essential. We saw it as the key to building the best product for our users. We needed to build trust with our users before they felt comfortable sharing some of their most personal and sensitive questions. Focusing on building trust demanded that we remove any extraneous content or processes, in order to provide our users with the experience that they needed and deserved.

* The Springster Privacy Policy, including information on the chatbot, is available here: http://us.hayspringster.com/footer-pages/springster-privacy-policy/
We learned that chatbots represent a wholesale shift in how to engage users, with our users demanding personable, useful, responsive, and always-on content. Content is a central aspect of chatbot development, and has significant multipliers. Great content drives user engagement, participation, and may lead to achieving major objectives such as behaviour change. In contrast, bad content creates frustration and, for our users, fails to provide them with the information that they need to navigate potentially life-changing situations.

Insight two: Chatbot writing is unique
Understanding the value of content

We found an interesting dichotomy particularly relevant to chatbots – the distinction between content and channel, and the interplay between these two elements. In other products and services, content is the primary way to achieve a desired impact, for example behaviour change, whereas the channel is the route to engage users and to deliver content. However, when developing our chatbot we found the distinction to be much more blurred. Here, the channel was also essential in building user knowledge and confidence over time. There was also a learning curve amongst some of our users, associated with the chatbot as discussed above. Some of our users were unfamiliar with receiving content through this medium, and interacting with a chatbot. We had to write content with this in mind.

One of the trickiest aspects of content development was linked to the always-on nature of the chatbot. It’s a unique channel to engage users as it’s not transactional like a website, where information is presented and read. Instead, it’s a continuous process, and it’s important to build an overall narrative in the direction of longer-term outcomes or behaviour change. This longer-term aspect is about building trust and a relationship with users. For Girl Effect, future engagement may include asking about topics of interest, and nudging users to make smaller behaviour changes - for example, affirmation of their choices and positive reinforcement.

This potential for continuous engagement also meant we had to ensure each piece of content was useful. Any superfluous content would have encouraged users to leave the chatbot, and discouraged exploration. The initial test phase of the chatbot highlighted more than 1,000 unique user journeys as users explored the chatbot (see Figure 7). This range of user engagement validated our work to make interaction with the chatbot, and content, as simple as possible.

Figure 7. The chatbot enabled tailored experiences resulting in 1,165 unique user journeys, but the most common journeys were influenced by chatbot structure and suggestions.
Build a content process

We found that building a streamlined process to develop content was essential. This included standardising content development by building a content wireframe featuring a template for each content aspect - whether quiz, engagement message, or core content. More widely, only one team member was responsible for uploading and managing content in RapidPro in order to prevent duplication. Naming conventions for content, and measurement, were also hugely important - particularly in order to ensure that key measurement points weren’t missed.

We drew on best practice from the private sector in developing our content approach. We rarely sent messages longer than 90 characters and limited consecutive messages to three. We also aimed to vary the character length between each message. Messages were categorised into types of content: introductions, main content, topic definitions, signposting options, and closing content. Similarly, each content type had its own formulation. For example when writing the main content, the writer would provide two messages of concrete information, an engagement prompt to move the conversation forward, and then an acknowledgement message to recognise the girl’s voice and to continue the discussion. We aimed for no more than three messages without having some interaction with the user - both to minimise users having to scroll due to small device screens, an issue highlighted during initial user workshops, and to drive engagement.

The testing phases explored the depth of conversations, and how best to provide ‘natural’ dialogue with users. Overall, consistency in the experience was key - which is why high-quality content is essential. However, content development was time-intensive and we learned to prepare additional content ahead of the second Sprint so as to focus our efforts during that week. Building the above process to enable continuous content creation was an enormous asset during the development stages, and resulted in useful content (Figure 8).

Figure 8. Girls overwhelmingly reported all message sets to be helpful with no content receiving less than 78% “helpful”
Humans are still important!

Writing content for a chatbot is a very specific skill, demanding a writer familiar with this way of communicating. We benefited from a full-time in-house writer, well-versed in chatbots. This also ensured that the chatbot had a consistent voice. The ‘always-on’ aspect impacted on areas beyond content, too. As our chatbot focused on providing answers to girls’ sensitive questions, we needed to be able to provide answers immediately - or with the illusion of immediacy. This meant signposting girls to other content, and providing holding responses until one of our team could respond. This demanded a team with extensive knowledge of reliable sources of information for girls, and encyclopedic knowledge of the information available on our other channels, as well as awareness of vetted services to which we could signpost girls.

However, the expectation of constant availability, in this setting, also had a more serious side. Due to the accessibility of the chatbot, Girl Effect had to ensure we had moderator cover to handle any real and significant disclosure and safeguarding issues. During the test phase, we had 243 triggers, with 61 being interactions that we would classify as a safeguarding alert. These interactions would be addressed via human intervention to connect girls with suitable support services.

We learned a lot about how best to support girls. In particular, finding the ‘right’ trigger words, so as to reduce false positives - girls who actually didn’t need help - was difficult. We erred on the side of caution here in order to be able to help as many users as possible. However, only 25% of total triggers were safeguarding alerts correctly flagged as a girl at risk - with the other 75% of alerts triggered by low-risk messages that did not require costly human resource. In some instances the safeguarding messages and automated message flows were also too simplified and, at times, resulted in inappropriate, unhelpful, or jarring responses.
Although many of our projects and programmes have a digital focus, we learned a lot from developing the chatbot — including incorporating best practice from the digital and tech sectors. Using Sprints, incorporating regular reporting, building cross-disciplinary team working, and embedding a focus on iteration and improvement, with the user experience at the centre, were key drivers.
Using Sprints to build a Minimum Viable Product

A Sprint is a time-limited period of development, in which a defined product is built. This product is often known as a Minimum Viable Product – an MVP, or prototype. Sprints are used heavily in the digital world, and we borrowed this process when building the chatbot. We put together a multi-disciplinary team – including content experts, developers, and measurement specialists; each with clear roles and responsibilities – to build a prototype. In just over two weeks, we went from an idea to a functioning chatbot on Facebook Messenger - and then began testing with a real audience.

Focusing on building an MVP highlighted a range of wider considerations. First, we had to be comfortable not building a finished product. The process was about testing, and learning, and recognising that failure wasn’t the end of something, and ensuring that measurement and evidence drive actionable decision-making and meaningful improvement. An MVP was about understanding what does and doesn’t work, often at significant pace.

Second, building an MVP demanded that we engage with some more strategic questions. As discussed above, content development is enormously time-intensive and we perhaps had not understood this prior to the first Sprint. With that in mind, we had to decide which content to prioritise now – in order to prove the chatbot concept – and which could come later in the development pathway. We had to focus on building the foundations first.

Ahead of the Sprint, we identified the best channel for building an MVP. We needed a channel that delivered the user experience required, achieved our outcomes in the most effective way, and met our wider requirements, including data privacy and security. When building an MVP, unless it’s a very specific proof-of-concept, we learned that it’s important to build on the platforms where your users are. With a limited timeframe, we also had to be proportionate in the amount of work and testing required, scope the product well, and be realistic with what we could launch, including managing our ambitions.
We selected Facebook Messenger, as the barriers to entry to developing a functioning chatbot are minimal - including seamless recruitment through Facebook adverts. It was also where many of our users were, where we could build something quickly, and it met our privacy requirements. Notably, at the time, other platforms did not allow masking of user phone numbers. Multi-channel may be the eventual goal for the chatbot, but the priority during the prototyping phase was to get something built and launched. However, we recognised that designing and building with scale in mind was important.

We wanted to build a ‘global’ prototype, but due to resources we focused on English-speaking communities around the world. We were alert to this limitation, though, and will be exploring translated content in a future testing phase. We also had to analyse data with this in mind and we expected less engagement in the test phase countries where English wasn’t as dominant a language as in other countries, although, this wasn’t always the case. Digital literacy was a factor across countries, though, and many users were unfamiliar with chatbots, Facebook Messenger, and communicating online.

We also had to be comfortable with other limitations of the prototype. Notably, users were likely to exhaust the content and the user experience wouldn’t be as slick as a fully-featured product. We also recognised that we would have to increase our reach to vulnerable girls as we developed the chatbot further. However, for us, testing was about proving the concept, and showcasing this to funders in order to scale the chatbot up with additional partnerships.
Rethinking success

We quickly learned that identifying the success of a chatbot is complex. We had to consider how to define positive outcomes for users, as well as validate the technology and its suitability for this audience. This challenge was complicated by the difficulties associated with conceptualising the outcomes of a chatbot, identifying any real-world impact, and validating, or applying rigour to, self-reported outcomes.

We used the testing phase to explore a few ways to consider success. First, was the role of quizzes in the chatbot - in order to see any initial changes in knowledge shifts amongst girls. Each topic had an associated quiz, and these were divided into baseline and endline quizzes. If a user entered the chatbot and went straight to a quiz, their performance in the quiz was measured as their baseline. After reading the content and attempting the quiz again, this second score was the endline measurement. We were careful not to position this approach as a strong marker of knowledge improvement. Instead, it was an initial exploration of how to explore the potential success of the chatbot.

We compared the average quiz score of users who completed both of these quizzes and found that, on aggregate, girls improved their understanding of topics by 36% on average after reading content (Figure 9). The improvement in knowledge across all topic areas was also statistically significant. This process allowed us to identify areas where users had a lower-than-expected initial knowledge of a topic, and prioritise them for future content development.

![Image](image.png)

**Figure 9. Increase in knowledge across topics, by average test score out of three (total n=419)**

*P<0.0001*
Secondly, we developed a series of outcomes that we could use to broadly evaluate the very initial success of the chatbot with this audience. These were largely utility focused, but self-reported and only measurable with a small number of engaged users. Overall, the chatbot raised girls’ awareness of contraception and confirmed their understanding of the information provided. However, there were mixed results in our ability to impact upon knowledge and attitudes. With regard to improving girls’ awareness of and confidence in asserting their rights, those self-reporting noted a mid-to-high level of having the confidence to act to improve their position in a relationship, or other emotional, situation.

Although these initial explorations of outcomes were promising, we learned that measuring the overall success of the chatbot is a little way off - and may require some changes to the product. This includes randomising the content selection menu, not currently possible in RapidPro, to understand whether content selection is driven by convenience or interest.

We also need to identify additional opportunities to explore changes in knowledge and understanding, beyond capturing the initial knowledge of users.

Overall, a greater volume of users and therefore data should allow us to explore differences between knowledge and attitudes across users, and then use these insights to guide new content development and structural changes.

In addition, the chatbot cannot currently determine behaviour change in our users. In order to do this, we need to build quizzes, notifications, and other interventions with return users in mind. This may allow us to explore behaviour change - for example: ‘Did you do x since we last chatted?’.

Future development may also review technical strategies with impact in mind, including potentially using check-ins at specific locations in the user journey. We will also need to consider the role of on-the-ground and service delivery partnerships, to understand the outcomes of the chatbot on real-world behaviour change. This will allow us to explore the entire user journey: from increasing knowledge and information using online support, to seeking offline services and guidance to improve her life.
However, beyond fixing minor issues, a central aspect of our post-launch work focused on retention and engagement - or re-engagement as many users hadn’t returned to the chatbot due to an initially sub-par experience, or for other reasons. As discussed earlier, we had a substantial number of users who experienced difficulty accessing content. In response to this, following the first testing phase, we made a significant number of changes to the functionality of the chatbot in order to improve the overall user experience.

We then ran a second phase of testing to consider the impact of these optimisations - as well as to initially explore how to re-engage users. As a team, we extensively discussed re-engagement and how best to do this. We were particularly concerned with biasing the process through incorporating incentives, including pushing particular content or ‘prizes’ such as quiz completion badges. We didn’t want users to exhaust particular content just to get a gift, as this would likely skew participation rates and weaken the insights we could gain. This further highlighted the complexity of effectively understanding user engagement.

Based on these discussions, we setup a two-stage re-engagement test. The first stage comprised of sending a notification message to the girls who had a sub-par experience, highlighting the improvements that we had made. We also sent a ‘welcome’ message from Big Sis, to users who had previously only interacted with Springbot. The second stage of the re-engagement test comprised of six re-engagement notification messages - split into three categories: a ‘true or false’ message, to prompt users to further explore quizzes; a ‘nice message’ without a direct call-to-action, for example a motivational quote; and a message promoting a new content topic. Each of these three groups had two different message variants: a total of six messages.

We randomised users - based on their persona selection, country, and previous experience with the chatbot - into one of these six groups. Each group was then sent one of the six messages, on a Saturday at the exact time at which they first initially engaged with the chatbot. This timing was an attempt to reduce any confounding associated with the timing of the message, although unfortunately we were not able to send messages on the day and time of original engagement. The following week, users were randomised to see a message from one of the remaining two groups, and the following week users would receive the final of the three message types that they had not yet received. An outline of the re-engagement process is at Figure 11.
As an aside, we extensively considered the gap between sending messages - so as not to contribute to any carry-over effect\(^*\). There is very little evidence or guidance around this effect in a digital setting, particularly in relation to notification messages. In addition, any guidance is completely non-existent in a lower-income setting - where users may only have intermittent access to phones\(^9\). We therefore decided that a gap of a week between each notification message should be sufficient.

More widely, we were also constrained by our ability to not randomise according to best practice, due to the lack of known characteristics of each user, as noted earlier. However, in response to this, we prioritised learning the impact of each message type. Therefore, users were randomised at the outset and then for each message, but not message type. With ‘proper’ randomisation, each week users would have had an equal probability of receiving one of the three message types, meaning some users may not have seen all three types and some users may have seen one type more than once. In the setup we implemented, users would see all three message types.

During this phase, we wanted to see if re-engagement was successful, and whether particular re-engagement messages could drive engagement with content or quizzes. Broadly, re-engagement was successful (Figure 12). We saw a number of users re-engage with the chatbot following receipt of a message. The message click-rate, the percentage of users clicking the message after receiving it, varied between 23% and 43%. Similarly, messages with quiz functionality (‘True or False?’) encouraged users to take the most quizzes, whilst those highlighting new content (‘New message set!’) were responsible for the highest number of topics started by users.

However, we were only able to draw limited insights from this testing due to unforeseen technical difficulties. In particular, we were not initially aware of limitations to the number of messages that can be sent to a user if they have not interacted with the chatbot for a length of time. This meant that some users only received part of a notification message, whilst many others received no message at all. We therefore ended this test one week early - and with 5% of messages being received by users. Similarly, the ‘Big Sis welcome’ message was not sent out due to technical difficulties merging the Big Sis and Springbot pages.

\(^*\) A carry-over effect is when the effects of one intervention (here, a message) impact on a successive intervention (in this example, a following message). If a carry-over effect is present, results could be biased as the impact of the second (or successive) message may be affected by those that were received beforehand.

\(^9\) Girl Effect’s research in partnership with the Vodafone Foundation highlighted the wide definition of mobile phone ‘access’ amongst adolescent girls. This research is available here: https://www.girlsandmobile.org
Insight four: Changing how we measure

Digital audiences are savvy and demanding. This context, combined with the always-on aspect of the channel and the centrality of the user experience, required that we develop a different approach to measurement. This included a combination of back-end metrics and those using more typical surveys. The former were substantially preferred by users. For the test phases, our measurement focused on utility: what was working for users and what wasn’t; and how to engage the most users, including conversion rates, the success of different acquisition channels, and how to retain users in the chatbot. These metrics will continue to be essential going forward.
Making measurement strategic

Measurement priorities changed depending on the phase of the chatbot project - whether it was the development, testing, or launch phases. They will also likely evolve as we look to scale-up the chatbot. This required a more strategic approach to measurement, including several trade-offs. For example, to date we focused on measuring immediate utility and generating actionable insights for iteration. Over the longer-term, though, measurement priorities will need to shift to explore deeper outcomes associated with the chatbot, including impact.

No matter the phase, though, we learned that measurement had to be agile and actionable. This had implications for the types of data that we could collect. Quantitative data had particular value as it could be analysed quickly - particularly as we ensured that data was standardised wherever possible. Bounded qualitative insights added some detail, but we limited these to the testing phase, or discrete periods, due to the significant analytical effort they required. Measurement had to be about driving change, and we would find it hard to sustain extensive analytical efforts over the longer-term if these were the sole drivers of iteration.

As discussed earlier, the user’s trust in the chatbot is paramount - and this also had to be the foundation of our measurement approach. We had to build trust with users, and demonstrate value, before looking to engage from a measurement perspective. Formal surveys and questioning are extractive, requiring users to provide information or data. When attempting these in a chatbot, without providing value or building any relationship, users will leave. There may be a role for surveys, in particular Lean Surveys\textsuperscript{10}, but their placement in the user journey must be carefully considered.

In our chatbot, we did include formal surveys - but these were only served to returning users, alongside questions also added into the navigation flow so as not to prevent users from reading the content of interest. The completion rate of the survey was 42%, and we had an overall response rate of 11% - the latter broadly in line with average online survey response rates\textsuperscript{11}. Interestingly, we were also able to see the optimum amount of questions to ask this group. We saw a drop-off after the second and sixth questions (Figure 10) - broadly aligning with other evidence in this area\textsuperscript{12}.

However, these numbers are not indicative of the overall performance of the survey in such a dynamic channel. Although the response and completion rates are relatively decent, the survey invitation - triggered a set amount of days after a user first started interacting with the chatbot - almost perfectly correlated with a substantial increase in users blocking the chatbot. This reaffirmed the difficulties of formally measuring this population. It also highlights the consequences of adding a more extractive interaction in a pathway where a user is attempting to gain value. Girls wanted answers to their questions, they didn’t want to answer ours - at least not without a greater demonstration of the value of the chatbot.

Perhaps confirming this, after refining the targeting of the survey to focus just on users who had engaged further with the chatbot the block rate reduced significantly. However, this increased targeting presented clear risks of selection bias in the survey and we caveated survey results accordingly.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{response_rate.png}
\caption{Drop-off for each additional question asked to users in survey}
\end{figure}

\textsuperscript{10} Lean research in international development focuses on gaining the best data, with minimal burden to participants. Acumen and the MIT D-Lab are leading efforts in this area.
\textsuperscript{11}Online survey provider SurveyGizmo notes an average response rate of 10-15% for surveys conducted with external participants (i.e. respondents outside of your organisation) https://www.surveygizmo.com/resources/blog/survey-response-rates/
\textsuperscript{12}Acumen’s Innovations in Impact Measurement research recommends a maximum of seven or eight questions, noting significant drop-off in response rates after this number of questions.
The realities of measuring in a digital environment

Developing the chatbot highlighted a number of factors that are very different to measuring offline interventions. Above all, we ensured that measurement didn’t ‘get in the way’ of the user experience. The user experience is paramount, and can’t be sacrificed for shorter-term learning. We also knew that measurement may detract so much from the user experience that it may actively bias the insights generated. This includes potentially minimising the success of the chatbot by discouraging many users from deeper engagement. Although undoubtedly a challenge, this encouraged us to pursue a creative approach to measurement.

One particularly useful approach was building in multiple measurement points into the back-end. This was a particular advantage of working in a digital setting, as this type of measurement doesn’t require any formal input and can be measured over the longer-term and at very regular intervals, reflecting the ‘always-on’ aspect of the chatbot. This contrasts with offline interventions, where measurement is often premised on capturing data at a particular moment in time.

With the chatbot, we were able to measure, track, and report a range of exciting metrics, based on user engagement and without detracting from their experience, to understand success. These were also layered with other insights from the above standard measurement approaches. We learned the importance of not over-measuring! The option to incorporate back-end metrics is a relative luxury for measurement professionals who aren’t used to such a wealth of potential data points. Careful consideration of what needs to be measured was important, to make analysis simpler and more strategic.

We also noted that the cost of measuring additional users is also drastically lower than offline. Notably, the barriers to a bigger sample are reduced as the cost to reach individual users is minimal. For example, the difference between engaging 4,000 users compared to 1,000 is much lower than it would be offline. We therefore focused on user recruitment and retention, as we can generate stronger insights with larger numbers of users. In order to inform iteration, we needed as many users to engage as possible.

We found that increasing the number of engaged users helped to identify broad overall trends, and reduce the ‘noise’ of weaker individual data. This reaffirmed the importance of not overly focusing on individuals. We also saw a necessary trade-off between the individual and group, or global, levels.

Segmenting too much, or focusing excessively on individual data points, risked us missing or failing to understand the broader insights that will drive iteration and eventual impact. This could have even caused bias in the user-flow, as fixating on the experience of a certain group of users may have affected how we approached development for all users.

One particular challenge that we encountered was the lack of baseline characteristics and information available - including demographic data. Although we collected some data through Facebook and surveying, this was sporadic and largely self-reported and not verifiable. This challenge reiterated the need for us to explore what is possible with outcome measurement. Going forward, and as we gain more and, hopefully, retain users, this issue may be reduced as we will be able to build a greater understanding of those interacting with the chatbot.
Thinking broader with measurement

We leveraged the wider user journey in our measurement approach, particularly in order to measure and understand some of our core priorities. Exploring these within the chatbot may have negatively impacted on the overall user experience.

We were keen to understand which of the two chatbot personas, the two chatbot identities and personalities that we had developed was preferred by users. However, technical limitations prevented users being able to select the persona in Facebook Messenger and, as noted above, this process would likely have detracted from the user experience. Instead, we used Facebook adverts, the main recruiting channel for the testing phase, to measure this outcome using custom referral links. Adverts, split across personas, were served randomly to our potential user-base. We were able to see which persona was preferred by users and which led to greatest engagement. The data provided from advert analytics also offered a range of demographic and other information that was not captured elsewhere.

With regard to the adverts, we found it essential to draw in social media expertise here, reaffirming the importance of a multidisciplinary team. Effective social media advertising campaigns have a range of different variables, objectives, and factors. Performance can also be hard to predict as pricing and how adverts are served is based on auctions and third party processes. Notably, Facebook runs adverts slowly at first, to guard against offensive campaigns, and then publishes them more widely. Similarly, advert performance can be iterative, with Facebook learning about who does and doesn’t click on adverts and then improving targeting and promotion accordingly.

We ran this aspect of the project like a typical advertising campaign, with a range of different advert designs, and using Facebook tools to target specific groups. However we did encounter some limitations in targeting adverts to our specific audience. Firstly, we had a relatively high cost-per-user when recruiting through this channel. We attempted to reduce this cost during the second phase of testing, as discussed below.

In addition, we also had to be cautious to not damage the future prospects of the chatbot by recruiting too many users. We knew that, as a prototype, the experience would not be perfect, so we had to be cautious about recruiting large numbers of users. This was a particular challenge in some countries, as the population of adolescent girls active on Facebook was smaller than elsewhere. In these countries we therefore risked over-saturating girls with Facebook adverts, which would have seen us reach the entire potential market for the future chatbot. If these girls then engaged with a prototype product, and had a negative experience, we would have struggled to launch a more refined version later.

Although the adverts did recruit more than 2,500 users for testing purposes, they broadly underperformed as an acquisition strategy and were comparatively expensive. We are investigating other strategies to inform future campaigns, and refinements we made to improve performance are discussed further in the following section.

We also noted the opportunity to leverage other tools to assist in chatbot measurement. This included online and digital analytics tools - such as Dashbot.io, Facebook Analytics, and others - allowing a wide range of accessible data to inform decision making. We saw that drawing on a range of analytics tools is important as each measures key factors in slightly different ways. Some, such as Facebook Analytics, provided us with data that we couldn’t capture at scale elsewhere, such as demographic information.
Insight five: Technology can be complex

Developing a chatbot offered us enormous opportunities, including reaching potentially millions of users, as well as revolutionising how we measure and understand success. However, we also had to engage with a number of complexities associated with this channel.

We also think that it’s important to highlight that digital and technology are not silver bullets nor panaceas. They are tools, and potential opportunities to deliver, aggregate, or augment impact. There are significant expectations associated with the transformational potential of digital and technology. However, these need to be tempered by a careful and strategic focus on the value and role of these tools in project, programme, and organisational working.
It’s about more than the technology

Throughout this process, we needed to ensure that we were not being led by the technology. This included periodic reflection to ensure that this remained the case. We had to make sure that our decisions were not driven by technology, and to confirm that a chatbot was the correct approach to engage users and achieve impact.

We couldn’t just rely on technology for the entire user experience - particularly with the difficult topics that users were exploring. We also needed to focus on human intervention including moderation, incorporating qualitative feedback, and founding this project on in-person user engagement at the very outset. From the start, we explored the careful balance between user expectations, and the realities of the technology.

All of this had implications for timeframes and resources. We had to budget for full-time moderation during key periods, protect project management resource, and undertake training and recruitment of staff including for content development, digital measurement, and RapidPro expertise; as well as increasing the digital competencies across the team more broadly. The project, and team, also benefited hugely from expert input and guidance from a chatbot specialist. Developing the chatbot was a cross-organisation endeavour, drawing on different areas and teams across Girl Effect - as well as a range of internal and external expertise.

More strategically, we had to engage with the inherent bias of digital. Delivering any service or product through a digital channel, by definition, prevents those without access to digital from accessing it. This bias is also shaped by the type of channel. Our users who were accessing the chatbot in Facebook Messenger may have been different to those using Line, or a different service, and different again to those who were not online at all.

Engaging with this bias is difficult, and there’s likely no ‘right’ way forward. Instead, we had to interrogate the strategic value of a chatbot and the benefits that it could deliver. Similarly, this selection bias was an essential consideration when we evaluated the chatbot. Users who clicked on it, engaged with it, or even had access to it in general may have had particular characteristics that weren’t actually representative of our priority audience. It may actually be that those who didn’t, or couldn’t, engage with the chatbot were actually those who would have benefited most from what it could offer.
Don’t forget about offline

Although a chatbot is a digital product, we still gained enormous value in conducting offline research and engagement to validate - or unpack - our initial chatbot thoughts. This included the extensive user research that we conducted via a series of in-person workshops at the outset. These workshops provided us with invaluable guidance on how best to frame the chatbot, and the types of content that girls were looking for.

Similarly, we’re now in the process of exploring further in-person workshops to gain in-person user feedback - and to provide an opportunity for more detailed interrogation of the chatbot, with our target users. These workshops should allow us to identify what’s working, what isn’t, and what more girls need.

We also used a combination of both online and offline measurement best practice.

With regard to the latter, this included developing a measurement framework, which is often at the heart of monitoring and evaluating ‘real-world’ interventions. The measurement framework, as with all other aspects of the chatbot, was subject to regular review. Metrics were refined during and after the test phase. Notably, we reduced the frequency of data collection for many metrics after the first testing phase, as we were comfortable that we had collected sufficient data to ‘prove’ key elements of the chatbot.
Launch is not the end!

Building and releasing a chatbot is comparatively simple. We learned that it’s everything that comes after any launch that presents the real challenge. In particular, this includes the time required for further bug-fixing and product tweaking - particularly when large numbers of users start interacting with the chatbot. For us, this included fixing referral links, and reducing the amount of measurement points so as not to over-measure.

With any digital intervention, the product or service is never truly ‘finished’. As discussed earlier, refinements and improvements are always needed.

Beyond existing users, it’s also about moving towards scale and reaching new users. As an initial attempt to explore this, we ran further social media tests to reduce the cost of user acquisition. We adapted our Facebook advert targeting, and the user acquisition journey. In particular, we targeted adverts at girls who were more likely to have Facebook Messenger installed on their phone, and avoided sending girls outside of Facebook by focusing on girls that were already interacting with Springster on Facebook. Using this targeting, girls could be driven to access the chatbot by using the ‘Send Message’ button on the Facebook page, rather than following a link which had previously resulted in user drop-off.

The success of these tests was mixed. We significantly reduced the cost of user acquisition - by over 60% in some markets - but this improvement may have been due to a lower ceiling price in the Facebook advert marketplace. Our user drop-off rates also did not improve. However, this testing did highlight several important implications for future marketing efforts. This includes testing other digital and social growth strategies (including peer sharing), developing a robust marketing strategy that does not rely solely on social media adverts; and perhaps focusing on longer-term delivery of any adverts - instead of timing them to be served during peak periods of Facebook user traffic.

As the safety of our users is our highest priority, we also wanted to test ways to improve the moderation process. In particular, we wanted to reduce girls accidentally being flagged as requiring safeguarding support due to ineffective trigger words. We added an intermediate step to triage moderation based on risk-level and self-reported need, and reduced the number of trigger words in order to exclude those that yielded the greatest number of false positives. These improvements resulted in a reduction in inappropriate triggers from 74% in the first testing period, to 12% in the second testing period - saving staff resource, and improving the user experience.
Reflections and conclusion

Following the second phase of testing, the team came together to look back on the results of both test phases, to discuss what we had learned - and how we had moved from proof-of-concept to a functioning product; and to explore next steps.
Reflections

At the outset of the project, we set out to answer two big questions: is a chatbot globally relevant to girls that we want to reach; and is a chatbot a channel where girls can get accurate information on sensitive topics? We felt that we had broadly, and positively, answered these two questions - but were keen to make further improvements in future phases.

With regard to building a chatbot that is globally relevant, we saw that user acquisition was strong - but more successful in some countries than in others. Similarly, we had developed a chatbot persona that in all tested markets was largely preferred by girls. This is perhaps an initial step toward creating a globally relevant persona for these users. As noted earlier, we also encountered some difficulties in retention - but initial indications from the second test phase show that there is high potential for developing retention and re-engagement strategies to bring users back into the chatbot.

Further localisation - of language, channel, and content - was seen to be an important priority. In particular, the completion rates of content, although high across countries, highlighted that some topics were also more relevant to girls in some countries than others. The testing phases allowed us to hypothesise that this difference between countries may also be due to technical variations - including differences in the technology used by girls, and the popularity of Facebook Messenger across countries.

Turning to our second big question, we have made strong progress toward building a channel where girls can get accurate information on sensitive topics. In particular, 90% of users surveyed reported feeling either comfortable, or a little comfortable, sharing their questions with the chatbot. The vast majority of users found the content was helpful, and some wanted even further depth and detail. Reaffirming the importance of localisation, we also saw some variation across countries in the sensitivity of topics for users.

We also reflected further on the data and insights that have been the foundation of this paper. We noted differences in users - some requiring more information than others; and some users wanting to explore the chatbot, whereas others wanted specific answers at a moment-in-time (including in emergency situations). We also identified opportunities to improve users’ experience in entering the chatbot. This all further highlighted the importance of user-centred design and user-driven processes in understanding and meeting the varied needs of users. Overall, though, building with retention in mind was seen to be a particular priority.

More widely, we generated a long list of further learning. This included applying a content-mindset to every aspect of the chatbot (including surveys); the usefulness of a marketing lens for user-experience decisions, such as targeting and retention; and the importance of striking a balance between providing a broad range of content to attract users, but a sufficiently useful depth of content to meet the knowledge gaps of chatbot users.
We also had three process-focused reflections. First, we successfully tested two personas using an ‘A/B’ test approach but never applied an ‘A/B’ approach to the chatbot itself - for example, testing a completely different chatbot alongside the version that we developed. Second, we built an MVP but we could have reduced this even further to an absolutely minimum viable product. Third, we benefited from an expert team - but not a dedicated one. The chatbot was built alongside team members’ other roles and responsibilities. We could have moved through Sprints, and testing, even quicker with a fully-dedicated team.

Building on this progress, and learning, we also explored next steps - and how we move to an even more developed product that fits into the wider Springster ecosystem of digital services and products that meet the needs of girls. This included immediate next steps with the current product - collating learning, document tidying, and user maintenance - and medium-term priorities. This latter group focused around building partnerships, and developing the foundations for the second phase of the product. Notably, this includes identifying priority markets for growth and impact; exploring any necessary design changes to best meet the needs of girls; and unpacking our technology needs and priorities.

* A/B tests (or multi-variant tests) are the Randomised Controlled Trials of the digital world. They allow the testing of different versions of digital products - such as different web-pages, or varied content placement - to see which is the most successful in engaging or converting users.
Conclusion

Overall, we successfully proved that the chatbot concept was an effective and useful way of engaging with and supporting girls, and providing the answers, information, and knowledge that they needed to answer essential and delicate questions.

We learned a significant amount from building the chatbot. This included the importance of building for the user, and prioritising their needs and experience; the essential role of personable and engaging content; the need for human input and guidance from testing to moderation; and drawing on best practice from the technology sector. We recognised the importance of testing, learning, and iterating; and the need to rethink our approach to measurement in order to understand what is and isn’t working. Above all, we recognised that a chatbot isn’t just about the technology: it’s about the role and value of this channel in providing our users with the product that they deserve.

In addition, our chatbot was built to appeal to a target audience of a specific age, and with the vulnerability criteria at the centre of Girl Effect’s approach. However, we found that it was very difficult to identify this target audience within the users of the chatbot. Without extensive surveying of our users to understand their realities – a guaranteed way to dissuade engagement with the chatbot – we had to rely on limited questions embedded within the chat to further engage with our audience. We are now exploring digital proxies relating to vulnerability, particularly in order to better identify and meet the needs of these priority users in an online and digital setting.

We developed, piloted, iterated, and evaluated a chatbot - moving from conception to the end of this process within one year. We’re enormously proud of this journey, and very keen to continue exploring how chatbots can support girls in a very important time of their life. We are now seeking funding to build on what we learned with our MVP. This may include further exploring quizzes, incorporating increased AI so that users can get more answers on demand and not have to wait for human intervention (although Human-in-the-Loop² may be a natural next step), and understanding how to drive - and measure - behaviour change.

As an aside, we also learned a valuable lesson about the wider technology ecosystem. We developed the chatbot during a period of significant sector-wide data privacy changes and discussions. During this time, Facebook halted the integration of artificial intelligence with human input. Humans guide machines to learn and refine their thinking and approaches, for example through defining initial data inputs and refining algorithms. This integration of artificial and human makes machine learning, and wider artificial intelligence, much more accurate.

²Human-in-the-loop technology combines artificial intelligence with human input. Humans guide machines to learn and refine their thinking and approaches, for example through defining initial data inputs and refining algorithms. This integration of artificial and human makes machine learning, and wider artificial intelligence, much more accurate.
of many external platforms, including some chatbot applications. We saw how businesses and organisations that were reliant on these applications had these deactivated - thereby demonstrating the complexities of relying on a third party platform.

Similarly, our progress more widely was also sometimes dependent on Facebook action - notably merging the Big Sis and Springbot pages after we elected to move forward with a single, and preferred, persona. Thankfully, our timetable was fluid enough to respond to this. However, these factors may reaffirm the importance of a multi-channel solution, and the usefulness of separating the channel and the back-end.

Overall, we started this process with no real guidance around the potential of chatbots in encouraging knowledge and attitude change in international development. We found even less evidence regarding the role that they could play in supporting adolescent girls in lower-income settings. Reflecting this, we committed to documenting each aspect of the project from the outset- including the successes, nuances, and challenges. We sincerely hope that this paper provides some guidance to other organisations exploring this area. Through sharing the lessons that we learned, we aim to - in a small way - support how chatbots can be used to change the lives of those in need.
girleffect.org

Girl Effect
Ingeni Building,
17 Broadwick Street,
London W1F 0DE

Charity No.1141155