Developing and testing a health app evaluation framework for organisations to recommend the best health apps for consumers in the Australian setting

June 2022

Authors:

Sithara Wanni Arachchige Dona

Dieu Nguyen

Mary Rose Angeles

Paul Cooper

Natalie Winter

Mary Lou Chatterton

Anna Peeters

Martin Hensher

Deakin Health Economics, School of Health and Social Development
Institute for Health Transformation
Deakin University

INSTITUTE FOR HEALTH TRANSFORMATION

ACKNOWLEDGEMENT OF COUNTRY

The researchers would like to acknowledge the Wadawurrung and Wurundjeri peoples of the Kulin Nation as the Traditional Owners of the lands on which we live and work. We pay our respect to their Elders past, present, emerging, and future.

FUNDING ACKNOWLEDGEMENT

This project was funded by Medibank Better Health Foundation (Medibank BHF). Professor Anna Peeters from Deakin University was supported by an NHMRC investigator grant.

ACKNOWLEDGEMENT OF INVOLVEMENT IN THE PROJECT

The researchers would like to acknowledge the Medibank staff's time and effort in working with the Deakin staff to test and improve the draft framework and advertise the survey on the Medibank website. In addition, the researchers are thankful to the focus group and survey participants for their time and valuable input.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

D-HEAL FMHE	The Deakin Health E-technologies Assessment Lab Framework for Mobile Health Evaluation
IRR	Inter-rater reliability
ICC	Intraclass correlation coefficient
STATA	Statistical Software for data science
CALD	Culturally and Linguistically Diverse
TGA	Therapeutic Good Administration
RCT	Randomised Control Trial
UX	User Experience
W3C	World Wide Web Consortium
T&C	Terms and Conditions
USA	The United States of America

EXECUTIVE SUMMARY

The number of health apps available on mobile devices continues to grow exponentially. However, there is very little authoritative guidance for consumers and health organisations to identify which of the health apps currently available in the market can be used safely and beneficially.

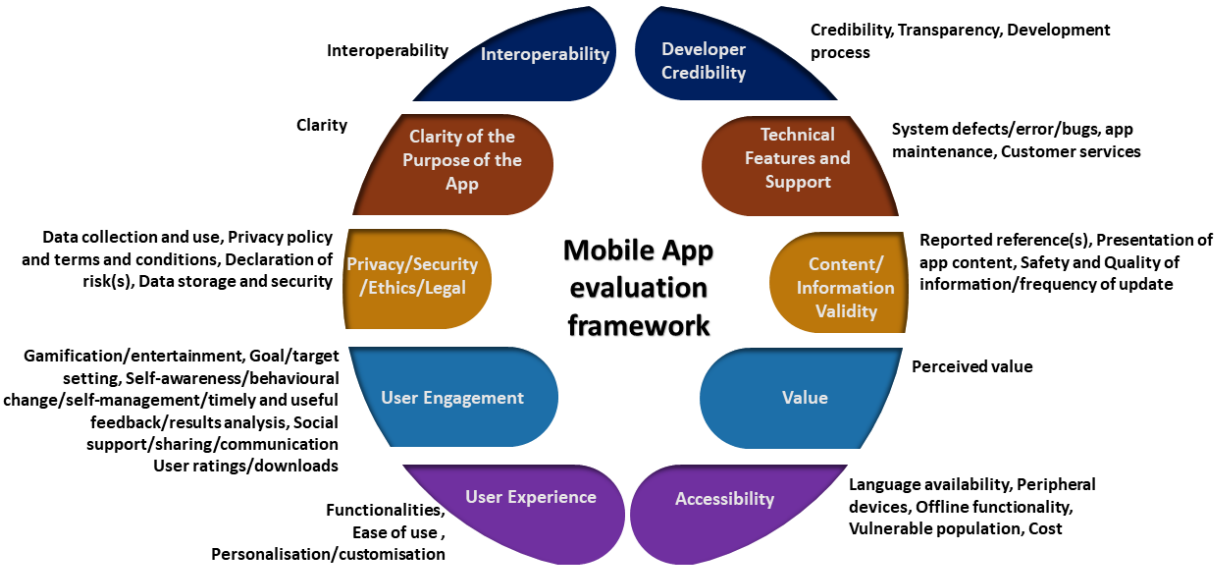
In this context, the aims of the project were:

1. Identify the priority domains from the scientific literature for evaluating digital health apps relevant to the Australian context.
2. Develop a framework based on the identified domains and domain items from the scientific literature.
3. Undertake health expert feedback on the identified domains and domain items' relevance, feasibility, and usability.
4. Evaluate the framework's usability when applied to apps covering general health issues and a range of common health issues or conditions frequently encountered in the general population.
5. Update the framework based on evaluation, expert and consumer feedback.

Phase I

Scoping Review: We conducted a scoping review of the literature to identify the priority domains most likely to be useful to include in a framework for evaluating and rating digital health apps. The review identified 97 studies that complied with a list of 430 assessment criteria. The most frequently used scaling mechanism was a 5-point Likert scale. Most studies have adopted summary statistics to generate the total scoring of each app, and the calculation of mean or average scores was the most popular approach. Some frameworks did not use any scaling or scoring mechanism and adopted criteria-based, pictorial, or descriptive approaches or "threshold" filter. Ten potential framework domains were identified across studies: *Clarity of Purpose of the App, Developer Credibility, Content/Information Validity, User Experience, User Engagement, Interoperability, Values, Technical Features and Support, Privacy/Security/Ethics/Legal, and Accessibility (Figure).* There are overlaps between some components of domains, suggesting there is some flexibility within frameworks and that categorisation of domains is not yet a standardised process. The selection of questions from the list of 430 assessment criteria, which was identified from our scoping review, for an app evaluation framework should be carefully conducted based on criteria including, but not limited to, the structure, the depth and the expected outcome from the question, and the subjectivity or objectivity, because individual perceptions of the quality of the questions can vary from one person to another. We published the scoping review in the Journal of the American Medical Informatics Association [1].

Domains identified from the scoping review



Focus Groups: We conducted four online focus group discussions with 17 health professionals and stakeholders with expertise in the digital health space regarding the relevance, feasibility, and usability of the identified domains from the scoping review. The majority of focus group participants (41%, n=7) were very familiar with a range of digital health apps on mobile devices, while only one participant (5%, n=1) had limited knowledge of a range of digital health apps on mobile devices Most of the focus group findings were similar to the scoping review findings. The other key findings, which we incorporated when drafting the framework, included the following.

The framework's perspective and the intended audience were deemed essential, especially when considering the domains' order or weight of importance. Regardless of who the intended audience is, *Privacy/Security/Ethical/Legal*, *Content/Information Validity*, and *Clarity of Purpose of the App* were felt to be the most important, while *Technical Features and Support* and *Interoperability* were seen as the least important. However, there was no definitive conclusion on the most critical domain.

For the *Accessibility* domain, health app evaluators should look beyond the language or online and offline functionality and review how the health apps extend or likely bridge the care for people with a disability or other vulnerable population groups.

On *Developer Credibility*, participants indicated that evaluators should consider the app maker's credibility and the factors driving the app content (i.e., source of funds or who commissioned the app). However, there were opposing views saying that the credibility of the maker should not be the basis of judging a quality app for the reason that credibility was easily manipulated or misleading.

User Engagement could be reviewed by checking the metrics or analytics of health apps, such as the number of downloads, prevalence of end-user visits, the frequency or constancy of usage, mobile traffic, and the attrition rate of end-users. Participants also recommended pilot testing

and a literature search on measuring *User Engagement,* because there are no standardised approaches to measure this domain.

The *Value* domain should be clearly defined as it varies depending on the perspective of the individual user, organisation, or authorities. One participant said that assessing all the other domains would mean that *Value* of the app has been assessed. Several domains, such as the *Clarity of the Purpose of the App, User Engagement, User Experience,* and *Accessibility* domains were thought to be connected to the *Value* domain. Some participants noted that the app's price is not an indicator of the app's value to users.

Development of the Framework: The development of the framework was initiated in Phase I, and findings from the scoping review were used to develop the framework by synthesizing common domains and items. As the COVID-19 pandemic emerged, the preliminary framework (Version 1) was first used to assess the best COVID-19 apps in Australia. The framework Version 1 was tested on COVID-19 apps weekly from 24th March to 13th May 2020, which were available for download and usable in Australia. At the final test of COVID-19 apps, we examined 16 apps after several screening steps against the preliminary framework domains and scored them according to the rating matrix. Based on our evaluation as of 20 May 2020, we recommended '*Coronavirus Australia'*, '*Healthdirect'*, '*My Aus COVID 19*', and '*COVID Safe'* apps as being the highest-scoring COVID-19 apps then available in Australia. The *Value* domain was seen as hard to assess.

Phase II

Assessment of the framework: We further developed the framework by integrating the findings from the scoping review, the focus groups, and the preliminary testing on COVID-19 apps. The resulting Deakin Health E-technologies Assessment Lab Framework for Mobile Health Evaluation (D-HEAL FMHE) Version 2 consisted of 27 questions with 5-point numerical rating scale responses to each question/item grouped into ten domains. Version 1 was improved by expansion from seven to ten domains while updating the 5-point scale from a rating matrix of 'good, indicative good, average, indicative poor, and poor' to a numerical 5-point rating scale with answer options tailored for each question.

We then evaluated mental health apps using this revised D-HEAL FMHE. Multiple evaluators from Deakin University and Medibank assessed the framework's usability. Apps designed for use by people with depression and anxiety/stress apps (and which were available on both the Apple and Google Play stores) were identified as the priority area of interest for evaluation using the framework. Three raters evaluated 49 apps to select the top 10 based on average total scores. In terms of resourcing, the average time duration for screening, data collection, and evaluation per app was 3 minutes (171 seconds), 5 minutes (326 seconds), and 1 hour respectively.

Inter-rater reliability testing: The inter-rater reliability among the three raters' responses was statistically analysed using Fleiss's Kappa and Intraclass correlation coefficient (ICC). In Fleiss's Kappa, none of the apps was at the extremes of either poor agreement (below 0) or almost perfect agreement (0.81-1.0). 25 of the 49 apps (51%) showed moderate agreement (0.41-0.6), followed by fair agreement (0.21-0.40) for 17 apps (35%). The calculated Fleiss's Kappa for each question of the Framework (three raters and all the apps subject to each question) showed that the most frequent level of agreement among questions (n=13/27) was "slight" agreement among three raters. This indicated a need to review and revise the questions in the framework, which we acted upon at a later stage of the framework development.

For the ICC, the pooled analysis of all the apps (n=49) showed a high consistency of observation amongst raters and apps. ICC analysis for individual apps indicated that 61% of rated apps have high consistency (reliability) in rating (ICC, Alpha > 0.7, n= 30). Subgroup analysis in each item and domain indicated the opportunity for improvement in questions 1-3, 9-13, 15, 17, 24, and 26. However, the issue of limited consistency could potentially be resolved by offering detailed training for questions 2-3, 13-15, and 17. Further refinement of questions 9-12 and 24-27 could also improve consistency. The domain *Privacy/Security/Ethics/Legal* consisted of questions 9-12 and had poorer average ICC in both poorly rated and highly rated apps. Due to the equivocal results for this domain, we will consider approaching this domain as an independent domain in our final version as a viable option to maintain the acceptable/good levels of consistency in other domains for the final scoring. At the same time, this domain remains a critical part of our framework.

Compared to Fleiss's Kappa analysis, the overall ICC result was consistent with some absolute agreement detected by app and item consistency rating tests. However, some apps were in disagreement and had a negative ICC score, suggesting sampling error attributed to a small sample size (i.e., low number of available raters).

Survey of the consumer experience of recommended apps: From the list of evaluated mental health apps, the top three scored apps ('ClearFear', 'Headspace', and 'Smiling Mind') were identified and communicated to the public by the Medibank website. Webpage visitors were invited to use these apps and participate in the survey study. We performed a pre-and post-survey of app users to collect consumers' experiences of using the top three mental health apps for depression and anxiety/stress, made available to the public by publishing them on the Medibank website along with a link to the survey. This survey aimed to investigate whether the consumer ratings on the apps correlated with the framework ratings on *User Engagement*, *User Experience* and *Value*. When they enrolled to participate in the study, participants were redirected to complete the baseline survey (S1). The first follow-up (S2) and the second follow-up (S3) were completed at the end of the first and fourth weeks, respectively.

The total number of survey participants was 55 at baseline, 27 at follow-up 1, and 23 at follow-up 3. Among those who had previously used mental health and wellbeing apps, the most frequently used app had been 'Smiling Mind' (n=15) followed by 'Headspace' (n=6) – both of which were also two of the three apps recommended by Medibank as part of this study.

Content/Information Validity was seen as the most important domain by all the participants and those who had previously used health apps, whereas *Interoperability* was ranked as the least important domain. This finding was consistent with the focus group findings. However, the rankings of those who had not previously used health apps differed significantly, with non-users ranking *Technical Features and Support* as the most important and *Developer Credibility* as the least important domains.

The mean ratings on the Likert scale for health app usability questions increased from S2 to S3. The change in the scale increased positively for 58% of the questions, which indicated that usability and satisfaction with the apps increased over time. All three apps had the same pattern of positive change. Respondents reported that their frequency of app use reduced overtime for all three apps. All the apps were rated as easy to use and easy to learn how to use, and navigation was convenient moving between screens. The Likert scale rating increased overtime for these features. App users were more satisfied with 'Smiling Mind' than 'Headspace', and our framework also rated these apps with a high score for user satisfaction related domains such as *User Engagement and User Experience*. 'Clear Fear' was the app ranked lowest for satisfaction by survey respondents, but the scoring from our framework indicated the opposite, where this

app scored high for satisfaction related domains. It should be noted that only two app users used and completed the surveys on 'Clear Fear', which could be the reason for opposite results on the 'Clear Fear' app. Interestingly, 'Headspace' and 'Clear Fear' users reported a reduction in their rating of the app over time as being beneficial for their health and well-being, but the change was not statistically significant.

Revising the framework: We then incorporated the evaluator feedback, inter-rater reliability, and survey findings to revise and refine the framework (Version 2). The majority of the questions were simplified or reworded to improve the agreement. The criteria within the questions were updated or reworded for ten questions. Answer options were updated or reworded where relevant. Sub-domains were updated. The *Value* domain was excluded. While the IRR result for the *Privacy/Security/Ethics/Legal* domain was problematic, the importance of the *Privacy/Security/Ethics/Legal* domain was one of the key findings from the literature review and the focus group discussions. However, we found it is the hardest and the weakest area of the framework to evaluate. Therefore, we assessed this domain separately without incorporating it into the overall scoring, but still keeping it as a domain in the framework. In the final step of Phase II, the framework was finalised with nine domains and 24 questions (Version 3).

Concluding the project: The framework initially synthesised ten domains for the assessment of health apps. These were primarily based on the evidence from our scoping review: *Clarity of Purpose of the App, Developer Credibility, Content/Information Validity, User Experience, User Engagement, Interoperability, Value*s, *Technical Features and Support, Privacy/Security/Ethics/Legal*, and *Accessibility*. The framework adopted the most commonly used 5-point Likert scale as a scaling mechanism and the total score for scoring. The average score of the total was used when multiple raters evaluated the same health app.

There are overlaps between some components of different domains, suggesting there is some flexibility within frameworks and that the categorisation of domains is not yet a standardised process. Therefore, we reviewed the ten domains based on our scoping review, the focus group study, and the evaluation process.

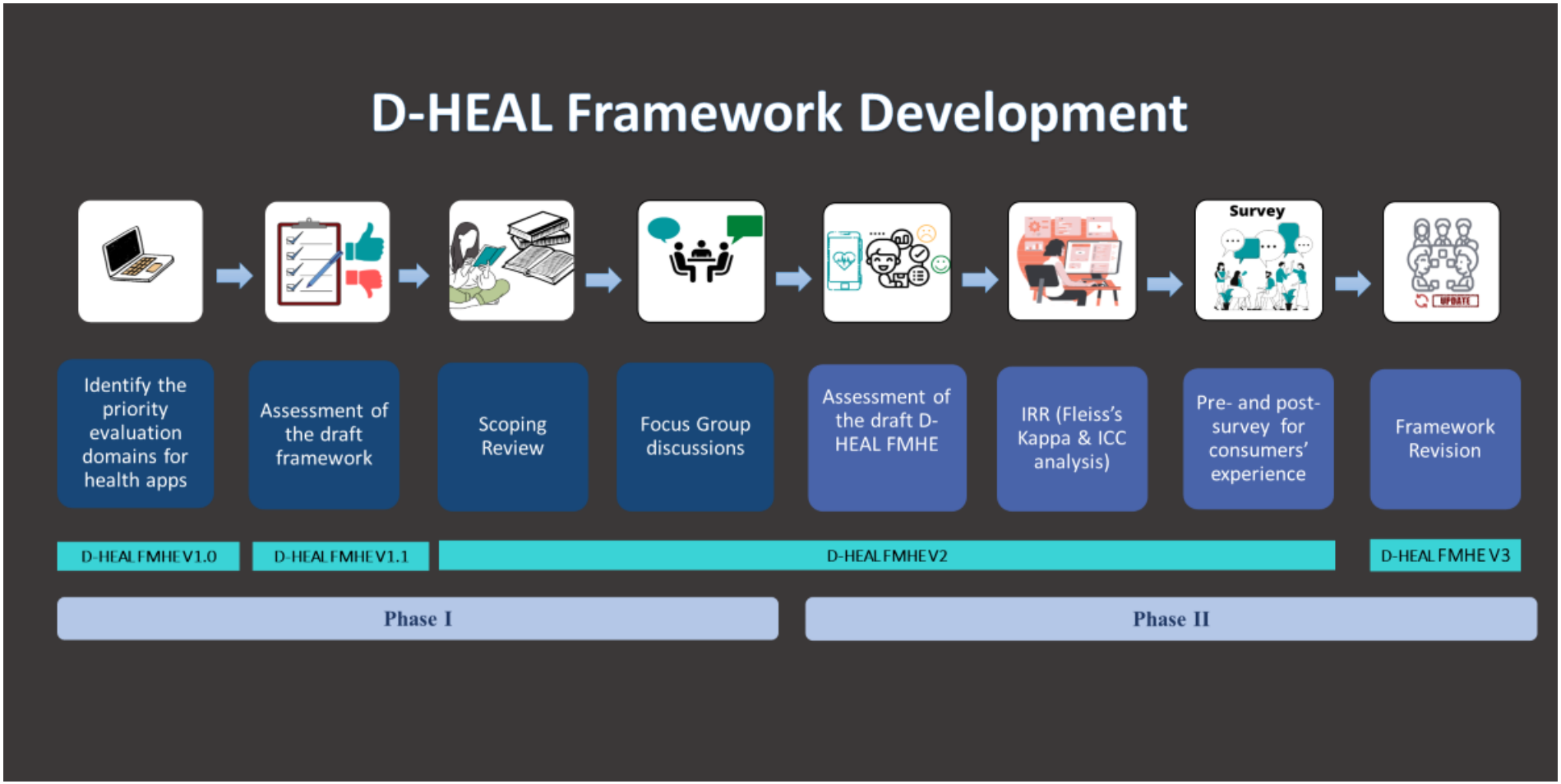
Value domain was deemed important based on the scoping review and focus group study. However, this domain, in which perceived value was a sub-domain, was determined to be a highly subjective domain to assess during the evaluation and was therefore removed from the draft framework (Version 3). Some focus group participants had suggested that carefully assessing all the other domains would mean that the "value" of the app has effectively been assessed. Similarly, the *Privacy/Security/Ethics/Legal* domain proved very complex to assess. Therefore, the scoring of this domain was not incorporated into the total scoring for health apps. Moreover, locating information to assess this domain needs thorough training. More research is needed to review the *Value* and *Privacy/Security/Ethics/Legal* domains.

Accessibility was identified as a new domain deemed crucial for assessment due to ethical and equity considerations. The benefit of *Accessibility* as a measurable domain arose in consideration of the framework on COVID-19 apps, where public criticism of the apps highlighted their poor support for CALD communities. In addition, the importance of *Accessibility* was highlighted in the focus group discussions. Our study found that in the mental health and wellness area, it is imperative for apps to support multiple languages and have simple user interfaces that have been ethnographically tested to ensure usability in a mixture of communities and people with disabilities.

Such aspects of fairness and equity are not frequently considered in other app evaluation frameworks, nor are they a feature of many major apps in the marketplace. Our report highlights the need for the requirement by the app stores to require clear evidence of accessibility support for apps specific to the health domain.

Finally, the D-HEAL framework was developed for organisations to evaluate health apps to promote the best health apps for their clients. Our iterative development and testing process has generated a framework which is suitable for use by organisations, and which provides ratings and recommendations which are consistent with the actual experiences of app users, yet which adds a level of technical assessment that most users will not be able to undertake themselves.

PICTORIAL PRESENTATION OF THE D-HEAL FMHE DEVELOPMENT



Domains


Content/Information Validity


Value


Clarity of Purpose of the App


Privacy/Security/Ethical/Legal


User Experience


Developer Credibility


Functionality

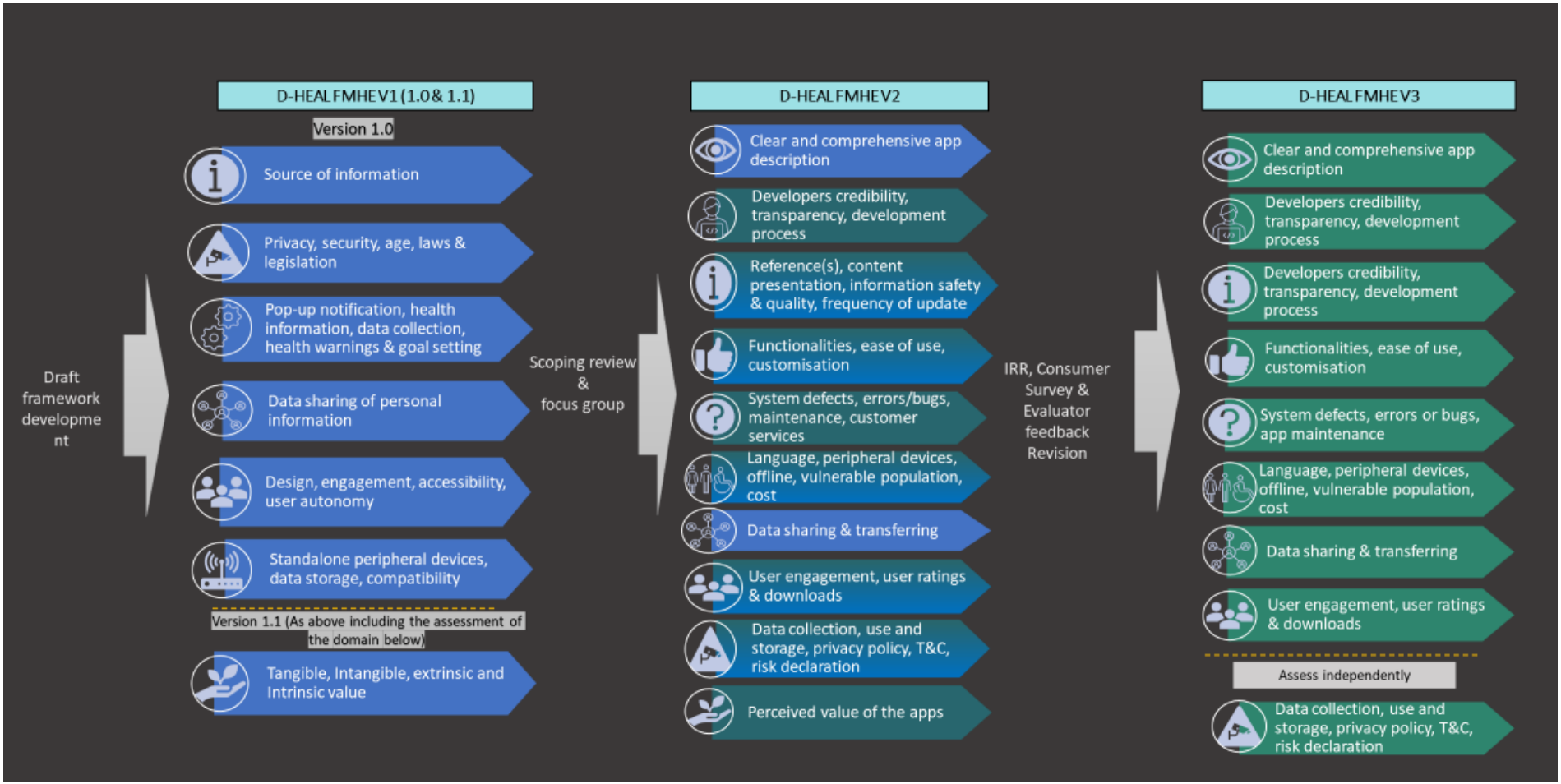

Technology Requirements


Technical Features and Support


User Engagement


Interoperability


Accessibility


D-HEALFMHEV1 (1.0& 1.1)
Version 1.0
Source of information
Privacy, security, age, laws & legislation
Pop-up notification, health information, data collection, health warnings & goal setting
Data sharing of personal information
Design, engagement, accessibility, user autonomy
Standalone peripheral devices, data storage, compatibility
Version 1.1 (As above including the assessment of the domain below)
Tangible, Intangible, extrinsic and Intrinsic value
Draft framework developme nt
Scoping review & focus group
D-HEALFMHEV2
Clear and comprehensive app description
Developers credibility, transparency, development process
Reference(s), content presentation, information safety & quality, frequency of update
Functionalities, ease of use, customisation
System defects, errors/bugs, maintenance, customer services
Language, peripheral devices, offline, vulnerable population, cost
Data sharing & transferring
User engagement, user ratings & downloads
Data collection, use and storage, privacy policy, T&C, risk declaration
Perceived value of the apps
IRR, Consumer Survey & Evaluator feedback Revision
D-HEALFMHEV3
Clear and comprehensive app description
Developers credibility, transparency, development process
Developers credibility, transparency, development process
Functionalities, ease of use, customisation
System defects, errors or bugs, app maintenance
Language, peripheral devices, offline, vulnerable population, cost
Data sharing & transferring
User engagement, user ratings & downloads
Assess independently
Data collection, use and storage, privacy policy, T&C, risk declaration

INTRODUCTION

Consumers currently have little guidance navigating the complex mobile health application (health apps) market. Digital health applications (be they for personal use or use by clinicians) have increased in recent years, with more than 325,000 mobile health apps available to download[2]. Recent press has highlighted the ineffectiveness of many apps related to health areas such as preconception/pregnancy and mental health[3]. There is very little authoritative guidance for consumers and health organisations to identify which health apps can be safe and beneficial to use in this market.

Globally, at present, no regulatory body adequately addresses the risk of using poorly designed health apps with clinical risks, including misinformation or poor outcomes. At the same time, consumers are overwhelmed by choice[4]. Approaches to the evaluation, accreditation, and regulation of health apps are developing. Regulators still tend to specifically consider those that might be regarded as 'medical apps' - in Australia, apps that represent "software as a medical device" are regulated by the Therapeutic Good Administration (TGA) and, for example, in the USA by the Food and Drug Administration (FDA). The characteristics of health apps in such regulation diverge significantly between countries[5]. In Australia, the TGA does not regulate health and lifestyle apps and those that do not meet the definition of a medical device under section 41BD of the *Therapeutic Good Act 1989* [6]. The Australian Commission on Health Care Safety & Quality developed National Safety and Quality Digital Mental Health Standards, which provide a framework for digital health services; however, regulation of individual apps is not provided [7]. Apart from these specific regulations in Australia, there is currently no systematic or broadly agreed approach to evaluating health apps for consumers' use that considers important elements such as usability, safety, privacy, and effectiveness.

Locally and internationally, government and private organisations have begun the work needed to evaluate digital health apps[8, 9]. While the need is increasingly recognised in Australia, only a few frameworks are available to assess health apps; however, these frameworks are not widely used by Australian health organisations to review and recommend apps to consumers due to various limitations. For example, the Mobile App Rating Scale (MARS) is a tool to evaluate lifestyle apps. However, this tool is not comprehensive due to the lack of attention to important aspects such as ethics, security and privacy[10]. At the same time, the Digital Health Guide is a framework designed for clinician use only[9].

Therefore, there is a need and an opportunity to evaluate the domains of a theoretical framework to enable the development of an Australian framework for rapid evaluation and rating of the health apps for Australian consumers in priority areas. Current examples of priority areas with a proliferation of apps and a demonstrated benefit to consumer engagement in their health include future pandemic readiness, preconception/pregnancy and mental health.

Deakin University identified this as a priority area and invested in supporting the theoretical work required to develop the Deakin Health E-technologies Assessment Lab (D-HEAL) and the proposed comprehensive generalised health app evaluation framework. Under a partnership between Deakin University and Medibank Private Limited (one of Australia's largest private health insurers) and the Medibank Better Health Foundation, this project was established to identify recommended health apps for mental health as a priority area for Medibank and enable consumer testing of the recommended health apps via the Medibank web portal to refine the framework.

The area of interest for our project was health apps, specifically focused on mental health. The definition of mental health was derived from the National Mental Health Plan 2003-2008,

which is “mental health is not only the absence of mental illness but is a state of emotional and social wellbeing in which the individual can cope with the normal stresses of life and achieve his or her potential”.[11] As the COVID-19 pandemic emerged, the preliminary framework was also used opportunistically to evaluate COVID-19 apps in the initial stage of the framework development.

Non-mobile native “web apps” (i.e., those that operate only via a browser) were excluded from this study which concentrated on native mobile apps available directly from the Australian Apple and Android App stores for use on mobiles and tablets. This focus on native apps is to screen at a first level those apps that have the assurance of the app store screening and privacy and security rules undertaken via Apple and Google app stores, and so that apps available within Australia could be specifically selected for evaluation

The main objectives of the project were to:

1. Identify the priority domains from the scientific literature for evaluating digital health apps relevant to the Australian context.

2. Develop a framework based on the identified domains and domain items from the scientific literature.

3. Undertake health expert feedback on the identified domains and domain items' relevance, feasibility, and usability.

4. Evaluate the framework's usability when applied to apps covering general health issues and a range of common health issues or conditions frequently encountered in the general population.

METHOD

The research project consisted of two phases:

The first phase included conducting a scoping review, developing a draft framework, and conducting focus group interviews with health experts.

The second phase involved assessing mental health apps using the draft framework, testing inter-rater reliability, and assessing the reliability of the framework rating compared to app users’ perceptions collected using validated survey instruments.

Phase I

Scoping review

We conducted a scoping review to identify the priority domains required to be included in a framework for evaluating and rating digital health apps.

The objectives of the scoping review of the literature on app evaluation were:

1. to identify the commonly recommended domains and gaps or deficits in existing frameworks for rating health apps
2. to analyse the advantages and disadvantages of the different questions/items categories used for evaluating the health app within each domain of the evaluation framework
3. to explore the advantages and disadvantages of existing health app evaluation frameworks’ scoring and evaluation methods aimed at individuals or mixed users (individuals and clinicians) and to provide the rationale behind the decisions.

We conducted the search across five databases (Medline Complete, CINAHL Complete, PubMed, Embase, and Scopus). Google and Google Scholar were used for grey literature search. Both Database and grey literature searches were limited from January 2018 to April 2020. Reference lists of identified systematic and narrative reviews from database search were screened for further eligible studies. No limitation was applied for the publication year for this backward searching. Therefore, overall, the time frame covered by this review was from 2011 to April 2020. The search terms were synonyms of "health apps," "evaluation," and "frameworks". Studies were included if they were related to health app evaluation frameworks, scaling or scoring, or evaluation mechanisms applied in these frameworks. These evaluation frameworks were about health apps for the general population or mixed users (clinicians and the public). No restriction was applied to study design, disease area(s), or age group. We excluded studies reported on health apps used only by a clinician(s), abstracts, incomplete or ongoing studies, posters, and studies with no full text.

The title and abstract of all the articles were divided into two groups and screened independently by two reviewers (S.W.A.D. and M.R.A.) using EndNote software X9.3.3. These were then divided equally between the same reviewers for full-text screening using the Rayyan platform for review management. A third reviewer (D.N.) checked excluded and included articles. Any disagreement was discussed with other authors (P.C. and M.H.) to reach a consensus. Two reviewers completed data extraction (S.W.A.D. and M.R.A.) independently and verified by a third reviewer (D.N.).

Three reviewers (S.W.A.D., M.R.A., and D.N.) conducted a thematic analysis and synthesized the available data. The included assessment criteria with similar characteristics were grouped into domains. The domain names were adapted from a previously identified review [12]. Any discrepancies between the three reviewers were resolved through discussion with other reviewers (M.H., P.C., A.P., N.W., and M.L.C.). Scaling and scoring mechanisms used in the frameworks were also investigated and analysed.

The scoping review was published in a Q1 journal- Journal of the American Medical Informatics Association JAMIA in March 2021 [1]. Link to the journal article: https://academic.oup.com/jamia/article/28/6/1318/6205942

Preliminary draft framework

A preliminary framework (Version 1) was developed alongside the scoping review based on the synthesized domains and items of the recently published Henson's framework, Moshi et al's systematic review, and Nouri et al's classification [1,8,12].

The preliminary framework was first used and tested on COVID-19 apps. A search was conducted in Apple Store, Google Play Store, Google search engine, and Similarweb from 24 March and repeated weekly to 13 May 2020 to ensure that all likely COVID-19 related native mobile health apps were identified for evaluation. The key terms used for the search were "COVID" and "coronavirus".

All apps were filtered based on their intended purpose, availability and Accessibility in Australia, language, target audience, developer, and collaborator. A requirement for inclusion was that apps were explicitly labelled as COVID-19 support. In Step Two, the filtered apps were assessed to identify any apparent capabilities of artificial intelligence (AI) and machine learning (ML) algorithms to provide advice that required active medical/clinical input or oversight.

Then the apps were selected for final inclusion from the previous step against key domains in the preliminary framework (Version 1 (sub-Version 1.0 and 1.1)). This regular testing on COVID-19 apps helped craft the preliminary framework by reframing the criteria and adding the *Value* domain. Therefore, we named the framework that was used for the very first COVID-19 app testing sub-Version 1.0 and the framework that was further developed and used for the last COVID-19 app testing sub-Version 1.1

The apps screened on 24 March were tested using Version 1.0, which includes six domains (Functionality; *User Experience, Adherence and Engagement; Interoperability; Technology Requirements; App Validity; Ethics, Data Privacy, Legal and Legislation).*

The framework Version 1.1 that was used for the final COVID-19 app testing included Value domain apart from all the above domains. These sub-steps complied to form Version 1 with seven domains (see Appendix 1.1 for the framework Version 1)

Finally, the remaining apps were given an overall rating using a five-scale rating based on the overall score: good, indicative good, average, indicative poor and poor.

The Phase II section of this report provides a description of the further development of the framework methodology.

Focus group interviews

The focus group phase aimed to acquire feedback from health professionals and stakeholders with sufficient expertise or informed opinions (referred to as "health experts" in this document) regarding the relevance, feasibility, and usability of the identified domains from the scoping review. Therefore, this step aimed to answer the following objectives:

1. Identify factors for consideration to improve the preliminary D-HEAL evaluation framework.
2. Explore aspects related to evaluating previously identified domains from the scoping review.
3. Identify gaps or information that was not identified in the scoping review.

Focus group discussions were undertaken over four days- 06th, 7th, 12^{th,} and 15th of October 2020. Before the initiation of focus groups, we provided an executive summary of the scoping review and a copy of the COVID-19 framework to participants.

Data Collection and Participants

Ethics approval was granted by the Deakin Human Ethics Advisory Group, Faculty of Health (HEAG-H 123_2020) to undertake this research. The participants were recruited from the D-HEAL advisory group, principal investigators, and the Medibank collaborator network. An email invitation letter was sent to the participants who met the following criteria: over 18 years old, English speakers with appropriate expertise and experience (in the digital health space regarding the relevance, feasibility, and usability of the identified domains), and commitment to provide feedback and participate in the research topic. Seventeen participants signed and consented to join the focus group, and these participants were health experts across different disciplines in Australia.

Focus group Procedure

Four online focus groups of 45 minutes duration were conducted via Zoom by three researchers (P.C., N.H., D.N.) to further refine the framework domains and items. Each of the four focus groups consisted of 5, 5, 4, and 3 participants representing professionals and academics from various relevant disciplines, government, health insurer and consumer perspectives.

Participants represented a demographic mix of gender, age, and experience. The participants were randomly grouped based on their availability for allocated dates for the focus group with a blind selection process (i.e., participants were blinded to the name of other participants). The focus group discussions were semi-structured and were guided by pre-prepared questions related to domains and items presented in Table 1. Open-ended questions encouraged the participants to provide their thoughts about the research topics.

An online tool (Qualtrics) was used at the start of the online sessions to supplement and collect focus group responses during the structured questions. Four questions were read aloud by the lead researcher. These included asking participants about their familiarity with health apps, current status and recommendation of health apps regulation in Australia, and if they would identify and rank the top three most important domains identified from the scoping review out of the possible 10 domains available.

The online session of discussion and questions were undertaken for 45 minutes and were recorded using Zoom. D.N. then transcribed and deidentified the focus group discussion to ensure confidentiality. P.C. and N.H. verified all transcriptions post-sessions to facilitate accurate data analysis, and the transcriptions were sent back to the focus group participants for their approval. All recordings were destroyed after the process; however, a log was generated to record participants' name and their given codes for reidentification purposes (if required). The response data for online questions were generated from the Qualtrics platform.

Analysis and Interpretation

This analysis followed the six phases of conducting a thematic analysis recommended by Braun & Clark[13] using the NVivo 12 software. Thematic Analysis is defined as identifying, analysing, and reporting patterns within a qualitative data set[13]. A semantic analysis approach was undertaken in this research, aiming to present a more comprehensive account of one specific theme related to a particular question[13]. Two other researchers (S.W.A.D., M.R.S.A.) conducted the coding and theming of the data to facilitate a blinded analysis. The two researchers discussed any discrepancies and the principal investigator (M.H.) adjudicated any unresolved areas. The data from the online poll were analysed using simple descriptive-analytical measures, such as frequency, median, and weighted mean in Microsoft Excel.

Phase II

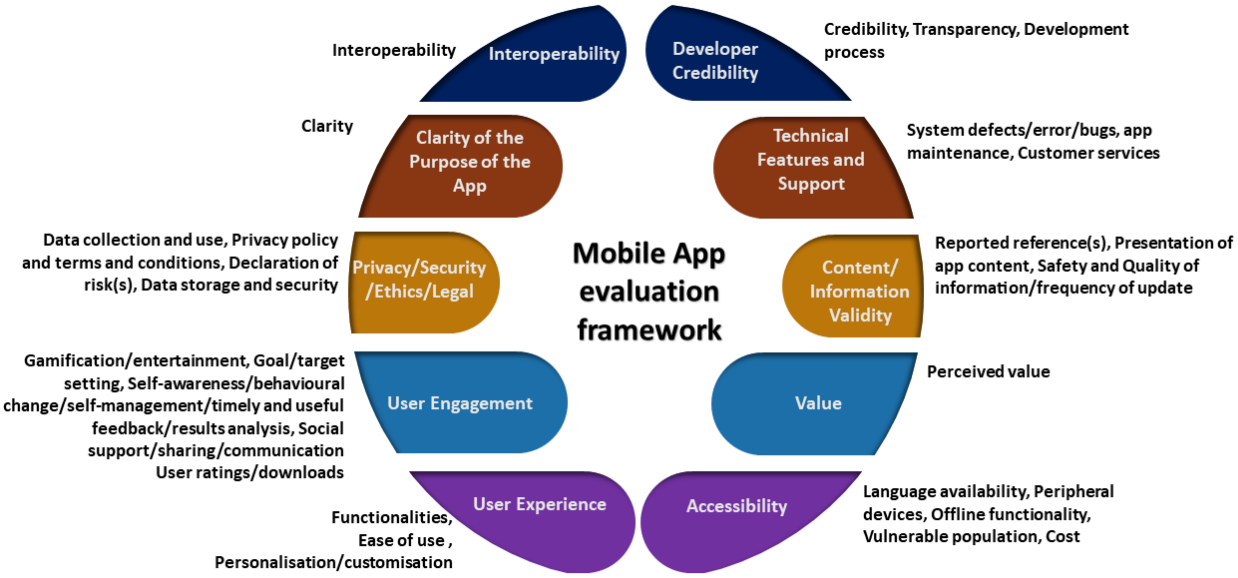
Assessment of the framework

The framework Version 1 with seven domains was further developed by integrating the findings from the scoping review and the focus groups, resulting in framework Version 2 (see Appendix 1.2_Framework Version 2). The research team discussed and analysed the domains and their sub-domains found from the scoping review and the data from the focus group interviews to obtain consensus on the main domains and sub-domains for the framework. The ‘question bank’ derived during the scoping review was then used as the source to select the most relevant questions based on the nature of domain and sub-domain given the feedback from the focus group interviews. Any discrepancies or disagreements were discussed with the advisory committee. The 5-point scaling was used for each question as it was the most frequently used scale across literature based on the scoping review.

As a result, this Deakin Health E-technologies Assessment Lab Framework for Mobile Health Evaluation (D-HEAL FMHE) Version 2 consists of ten domains with a 5-point scale in numerical *Value*s and 27 questions. The ten domains include *Clarity of Purpose of the App*, *Developer Credibility*, Content/Information Validity, *User Experience*, *User Engagement*,

Interoperability, *Value*, *Technical Features and Support*, Privacy/security/ethical/legal, and *Accessibility* (Figure 1).

Figure 1: Ten domains of the D-HEAL FMHE Version 2



The framework's usability was then assessed by multiple evaluators from Deakin University and Medibank using a set of mental health apps for depression and anxiety/stress identified by Deakin researchers. The selection process of the health apps is described below. We investigated the time to complete the framework, the source and ease of finding information, the confidence in the response and the final overall rating. The inter-rater reliability among the raters' responses on the evaluation was statistically analysed using Fleiss's Kappa on STATA and Intraclass correlation coefficient (ICC) on SPSS.

The final app selection process involved three steps: app search, app screening and app evaluation.

Search strategy

A search strategy was developed to identify mental health-related apps, as mental health is a key component of a person's overall health and wellbeing. It included depression, anxiety and stress, and general mental health using synonyms for these disease categories. The search terms were mental health, mindfulness, Selfcare, wellbeing, Meditat, sleep, Mind, Mood, CBT, Depression, Depress, depressed, depressive, anxiety, stress, fear, panic, and worry.

The search was conducted from 15th March 2021 to 19th March 2021 of the Apple app store and Google Play store (i.e., native phone apps). We considered only native apps available in both Apple and Google stores for this study. All apps found in one store were manually checked for availability in the other store to avoid search bias. All apps common to both stores were then screened. App name and the developer's name was recorded for all the apps to ensure correct identification since names sometimes varied between mobile device stores.

App screening

The following inclusion criteria were applied to screen out apps that were not suitable for Australian population usage using data from the app title and app description. If an app did not meet any of the following criteria, it was not included in the evaluation:

- App is related to the health area of interest (for this exercise, all depression, anxiety and general mental health apps).
- App supports the English language.
- App is usable in Australia (e.g., download and accessible without the requirement for the user to enter zip code or phone number from another country).
- App is available without prescription by health professionals (such as apps requiring access codes or passwords).
- App updates have been made available within the last 18 months.

During the screening, data were collected for all the screened apps. The screening process was conducted from 24th March 2021 to 11th April 2021. A final list of apps is suitable for further evaluation was generated. At the end of the screening, mental Health apps common to both Apple and Google Play were categorised based on the type of mental health disorder and the app's primary purpose as on the app description. The categories are as follows:

Categories:

- Cat 1 Depression
- Cat 2 Anxiety and Stress
- Cat 3 General mental health (i.e., meditation, mood tracker, etc.)
- Cat 4 Others (Wellbeing, Suicidal thoughts, Self-harm, borderline personality disorder, hypnotherapy)
- Cat 5 Mixed

Sub-categories (SUB-CAT):

- SUB-CAT 1: General risk reduction/disorder prevention (I.e., Apps that focus on both the prevention of general and disease-specific risk factors)
- SUB-CAT 2: Disorder management (I.e., Apps that focus on the management of symptoms, detection of disease and improvement of disease literacy, apps that use risk assessment methods)
- SUB-CAT 3: Care management (I.e., Apps that involve both clinicians and patients for the treatment management in an inpatient or outpatient care setting)
- SUB-CAT 4: Communication (I.e., Apps that allow the patient to connect with support groups and forums)

App evaluation

For this study, Apps for depression and or anxiety & stress were identified as the primary area of interest for evaluation (using the Framework).

Apps included for evaluation:

- Apps on general risk reduction/disorder prevention (i.e., Apps that focus on both the prevention of general and disease-specific risk factors)
- Apps on disorder management (i.e., Apps that focus on managing symptoms, detecting disease and improving disease literacy, or risk assessment).

- Apps updated within the last 12 months.

Apps excluded from the evaluation:

- Apps on SUB-CAT 3 care management (I.e., Apps that involve both clinicians and patients for the treatment management in an inpatient or outpatient care setting) and only
- Apps on SUB-CAT 4 communication (I.e., Apps that allow the patient to connect with support groups and forums).
- The target population- Some apps target a specific population. Apps were excluded on this basis as they specifically targeted the following groups: veterans, service members, workplace health, women, Athletes, children, college students, employees, real estate professionals in Australia & New Zealand, LGBTQ, Former police officers, police employees, and their families, students in Ireland, Men, college and university students in the UK, for professionals, child-serving professionals, and caregivers.
- Apps in Cat 3, 4 and 5 disease categories were excluded.
- Apps without the update history/information

We screened 5,522 mental health apps during the app search, with 3,023 apps in Google Play Store and 2,499 apps in the Apple App Store. After a manual duplicate check eliminating 1,794 apps, a total of 3,728 were reviewed for title screening. Five hundred eighty-nine apps available in both Stores were included for the next screening.

Finally, the 49 apps selected for evaluation in depression and anxiety/stress categories were assessed against the criteria D-HEAL FMHE Version 2 to choose the best ten apps rated the highest. The shortlisted 49 apps were processed through the full evaluation. The evaluation was conducted from 7th May 2021 to 13 May 2021 using the D-HEAL FMHE framework. Three raters evaluated each app. Based on the scoping review, it was advised to use the app for at least 10 minutes before evaluation. The instructions were provided for each question within the framework. The average time to complete evaluation for an app was one hour. The average total scoring for each app was calculated to find the top 10 apps.

Inter-rater reliability testing

Inter-rater reliability of app evaluation for the three raters was tested using Fleiss's Kappa and Intraclass correlation coefficient (ICC)

Fleiss's Kappa

Fleiss's Kappa for interrater reliability was conducted using STATA. Three raters rated all items of the framework for all the apps. The analysis was done in two pathways:

1. Per app: questions as subjects, interrater reliability was measured among three raters for each app using data for all the questions.
2. Per question: apps as subjects, interrater reliability was measured among three raters for each question using data for all the apps.

The STATA output gives kappa values for each outcome separately against the remaining outcomes. In our analysis, the outcomes were 1 to 5 rating scale. The combined value of Kappa is the appropriately weighted average of the individual kappa and is used as the interrater reliability[14].

Interpretation[14]: Fleiss's Kappa Value and agreement is as follows: < 0 poor agreement, 0.00-0.20 slight agreement, 0.21-0.40 fair agreement, 0.41-0.60 moderate agreement, 0.61-0.8 substantial agreement, and 0.81-1.0 almost perfect agreement.

Intraclass correlation coefficient (ICC/Cronbach Alpha)

ICC analysis/Cronbach Alpha was also conducted using SPSS to detect inter-rater reliability. Three raters rated all items and apps. Multiple analysis was taken, including:

1. Pool analysis: an overall agreement in all rated apps
2. Subgroup analysis: taking multiple approaches to analyse the consistency
    a. Approach 1: Per rated apps
    - Identified app with high score
    - Identified app with a poor score
    - Stratified app with a negative score to check all items as described in approach "per items"

    b. Approach 2: Per items
    - Identified the items with high consistency in rating
    - Verify the poor consistency item by stratifying highest/lowest Alpha scores to confirm that the items have reliability issues.

    c. Approach 3: Per domain and item(s) with domain
    - Identified domain has poor internal reliability/consistency
    - Identified items with low reliability and poor consistency

Please note that app/items with scores > 0.65 will be rounded up to 0.7 in the analysis.
Interpretation[15]: Alpha value and reliability in agreement is as follows: < 0.5 poor, 0.5- 0.75 moderate, 0.75- 0.9 good, and > 0.9 excellent.

As many guidelines indicate Alpha<0.5 as a cut off point for reliability, we also refer to ICC< 0.5 as unacceptable reliability amongst raters. Tavakol and Dennick 2011[16] noted a range of acceptable values (0.7-0.95), with Alpha > 0.9 recommended as the maximum Alpha value. The items which will be flagged as poor consistency will require further attention (either removal or improvement)

Survey of the consumer experience of recommended apps

A pre-and post-survey of app users was performed to collect consumers' experience of using the app, *User Engagement* and perceived value from the user's perspective.

The objectives of the survey were:

1. Measuring app *User Experience* of different apps (and correlation with Framework ratings)

2. Measuring app *User Engagement* with/use of apps – perceived and self-reported engagement.

3. Measuring perceived value to users of apps – perceptions of achievement of aims and value for money

The survey was developed and completed using the Deakin-standard Qualtrics online platform. The top three apps ('ClearFear', 'Headspace', 'Smiling Mind') out of the top 10 mental health apps for depression, anxiety and stress were communicated to the public on the Medibank website, along with a study description and the survey link.

The Medibank app recommendation page included a prominent invitation to opt-in to our research study at the point of selecting an app. Prospective participants were only invited into

the study if they decided to download an app to use. Participation and completion of the three surveys were encouraged and appreciated by allowing them to enter into a prize draw to win one of the electronic vouchers (e-voucher) worth AUD50 for 20 participants who fully completed all the surveys. The winners were randomly selected using the MS excel "Rand" formula to offer a fair and unbiased winner selection method. The prize draw mechanism is under the terms and conditions of Deakin University. A copy was given to the participants together with the PLS and consent form.

When members chose to enrol in the study, they were redirected to an online PLS and consent form. If they provided consent, they were redirected automatically to a form collecting their email address and assigning a "Random ID". The Qualtrics platform automatically generated a random ID enabling participants' anonymity. Subsequently, they were automatically redirected to complete the baseline survey (S1). An automatic e-mail attached with respective survey links were sent to the participants to complete the two follow-up questionnaires one week (S2) and four weeks (S3) post enrolment. Qualtrics automatically reminded the participants to complete the survey with a scheduled setting using the "E-mail trigger or Email task" function. Once the participants completed the S3 questionnaire, and at their discretion, they were redirected to another link (S4), allowing them to enter the raffle or opt-out. Only the contact form and S4 questionnaire dataset had the e-mail address, which allowed follow-up triggers, selection and communication of the winners for the prize draw. Moreover, these two datasets (containing the email addresses) were kept independently from the other project data, were not linked to survey responses, and were destroyed after winners claimed their prizes. Therefore, datasets for result analysis (S1, S2, S3) are unidentifiable to the researchers.

Quantitative data from the questionnaire were analysed using STATA/SPSS. Open-ended questions were thematically analysed using NVivo12. Recruitment ran until the end of November.

Concluding the research project

The inter-rater reliability and survey findings were then incorporated to revise and refine the framework. (see Appendix 1.2_Framework Version 3). Detailed information is in the "Framework Revision from the Workshop" section of this report.

RESULTS

The findings for each phase and step are shown in the following section.

Scoping review

Ninety-seven studies were included for the final synthesis. Based on analysis of the literature review, our research team identified ten proposed domains for evaluating health apps, as summarised in Table 1. Full details are provided in the published scoping review[1].

There are overlaps between some components of these domains, suggesting there is some flexibility within frameworks and that categorising domains is not yet a standardised process.

The selection of questions from the list for an app evaluation framework should be carefully conducted based on criteria including, but not limited to, the structure, the depth and the expected outcome from the question, and the subjectivity or objectivity, because individual perceptions on the quality of the questions can vary from one person to another. The most frequently reported scaling method was a point system: the 5-point Likert scale. The most

popular approach for scoring was the calculation of mean/average scores. Some studies have used different weightings of domains, while some frameworks did not use a scoring approach but used either a pyramid style model or a descriptive evaluation.

Table 1: Proposed domains for a general health apps evaluation framework

No	Proposed domains	Coverage/Definition
01	*Clarity of Purpose of the App*	A clear statement of the intended purpose of the app and the specificity of the users or the disease.
02	*Developer Credibility*	Transparency of the app development and testing process; and accountability and credibility of the app developer, funders, affiliations, and sponsors.
03	*Content/Information Validity*	Readability, credibility, characteristics, and quality of the information in the health app.
04	*User Experience* (UX)	The overall experience of using an app in terms of its user-friendliness, design features, and its ability to consider user preference through a personalization function.
05	*User Engagement*	The extent of how apps maintain user retention using functionalities such as gamification, forums, and the use of behaviour techniques, as well as the extent of social support
06	*Interoperability*	This refers to the data sharing and data transfer capabilities of the health apps.
07	*Value*	Perceived benefits of the apps and advantages associated with the use of health apps.
08	*Technical Features and Support*	Health apps that are free from defects, errors, bugs, and quantity and timely updates. Technical support and service quality provided within the app.
09	*Privacy/Security/Ethical/Legal*	Privacy and security domains pertain to data protection, cybersecurity, and encryption mechanisms for storage and data transmission. Legalities of the health apps look at whether the health apps adhere to guidelines and have disclaimers concerning clinical accountability.
10	*Accessibility*	This pertains to the ability of health apps to capture a wider audience and bridge the gap in access to health apps and healthcare services for vulnerable populations/people with disabilities.

Recommendations that came out from the scoping review for a new health evaluation framework are as follows:

- *User Experience* (UX), *User Engagement*, and social support should be evaluated separately. Disentangling *User Engagement* from the usability domain allows a wide array of questions under the usability domain. It allows a more straightforward evaluation of apps that attempt to enhance engagement and retention of users mainly by using behavioural techniques/principles.
- *Accessibility* was identified as a new domain deemed crucial for assessment due to ethical and equity considerations.
- Questions/items to assess an app should be clear, concise, specific, and objective. Our study has collated a library of specific questions and criteria from our reviewed studies.
- Due to a wide range of scaling mechanisms adopted across evaluation frameworks, we could not draw a conclusive recommendation on a single ‘best method’. However, health app assessments have adopted different approaches such as 3/4/5/7/10 point scales, dichotomous questions (answerable by yes/no or presence or absence option), or even open-ended questions deemed appropriate for the domain. These sets of methods can be used in combination. However, there are arguments from the literature [17, 18] that a simpler point scale would result in higher validity than more complex scales.
- One of our recommendations is to consider adapting the pyramid approach with a scoring mechanism. Currently, the available pyramid model framework does not have

a scoring system, and it is a framework with a hierarchical approach to evaluate apps. However, there is a possibility that a point system can be incorporated into the pyramid style to evaluate each level of the framework and provide a better presentation of results.

- The scoring for objective and subjective assessments should be calculated separately to limit assessment bias. In some frameworks, only objective domains have been scored.
- A user manual should also be developed to ensure that all assessors follow a guideline to evaluate apps.
- Prior to evaluation, evaluators should use apps for a minimum of 10 to 15 mins to ensure that apps have been explored adequately.

Further Recommendations and Considerations from the scoping review:

- Few studies have incorporated a *Value* domain in their frameworks; thus, more research is needed to review this category. However, some criteria noted in the question bank may be helpful to construct questions to evaluate this domain.
- A framework that addresses the limitations of currently available frameworks would allow for a consistent evaluation of different health apps for general users. This study will aid in the development of such a framework. Based on our review, we propose to develop a general evaluation framework.
- We recommend a tool comprised of valid and reliable questions, using a simple scoring method that can be used by the public and therefore does not require technical expertise to assess each of the domains in the general evaluation framework. Although not disease-specific, this framework could also support clinicians to filter apps based on the identified domains and thereby assist the consumer in choosing quality, valid, relevant, and valuable health apps.
- Most studies acknowledged a lack of validity and reliability testing and metrics that measure the level of engagement. Therefore, undertaking testing of any proposed evaluation framework would be important in addressing the aspect of *Content/Information Validity*.

Focus group data analysis

41% (7/17) of the focus group participants were very familiar with a range of digital health apps on mobile devices. 18% (3/17) were extremely familiar. 35% (2/17) were somewhat familiar, and only one participant had limited knowledge of mobile health apps.

Four main themes emerged during this analysis. The first theme concerned the need to regulate health apps in the Australian Market. The second theme provided overall recommendations or points for reflection when developing a general health apps evaluation framework. The third theme provided information on how to evaluate domains, while the last theme was about the relationship of domains and how one domain could not be disentangled from another. Appendix 2 summarises the key findings from the focus group data.

1. The need for health apps governance

Several participants raised concerns about health apps being unregulated in the Australian market. The majority of the participants (N=7, 41%) thought that there is currently only limited regulation of health apps in the Australian market. 29% (N=5/17) of the participants thought

there is no regulation of health apps in Australia, while 6% (N=1/17) noted that health apps are somewhat regulated. Four of the respondents (N=4/17) were not aware of health apps regulations in Australia. Most of the participants agreed that health apps should be regulated, and they would like to see health apps regulations modified or expanded in Australia. When they were asked what needs to be modified or expanded, participants indicated modifications or expansions in regulations, the accuracy of contents, and developing a curated set of health apps using an evaluation framework.

2. General points for framework

2.1 Audience for the framework

Some participants raised the significance of the perspective or the intended audience for the app evaluation framework. It was reported that when developing a general framework, the priority or the hierarchy of different domains would be different from an individual or organisational-level perspective.

2.2 The importance of domains

Privacy/Security/Ethical/Legal was ranked as the most important domain (frequency 7/17, median 2), followed by *Content/Information Validity* as the second (frequency 5/17, median 3) and the *Clarity of Purpose of the App* as the third (frequency 4/17, median 4). On the other hand, *Content/Information Validity* was rated as the most important domain based on the weighted mean (3), followed by *Privacy/Security/Ethical/Legal* (3.12). *Interoperability* and *Technical Features and Support* were consistently rated as the least important domains under both ranking systems. Appendices 2, 3, and 4 provide the median, frequency, and weighted mean for each domain. Figure 2 shows the order of the domains based on the median value of the ranking among 17 participants. A low to high median indicates the most important to the least important domain respectively. Figure 3 represents the order of the domains based on the weighted average of ranking. A low to high weighted mean indicates the most important to the least important domain respectively.

Figure 2: The order of the domains based on the median value of ranking

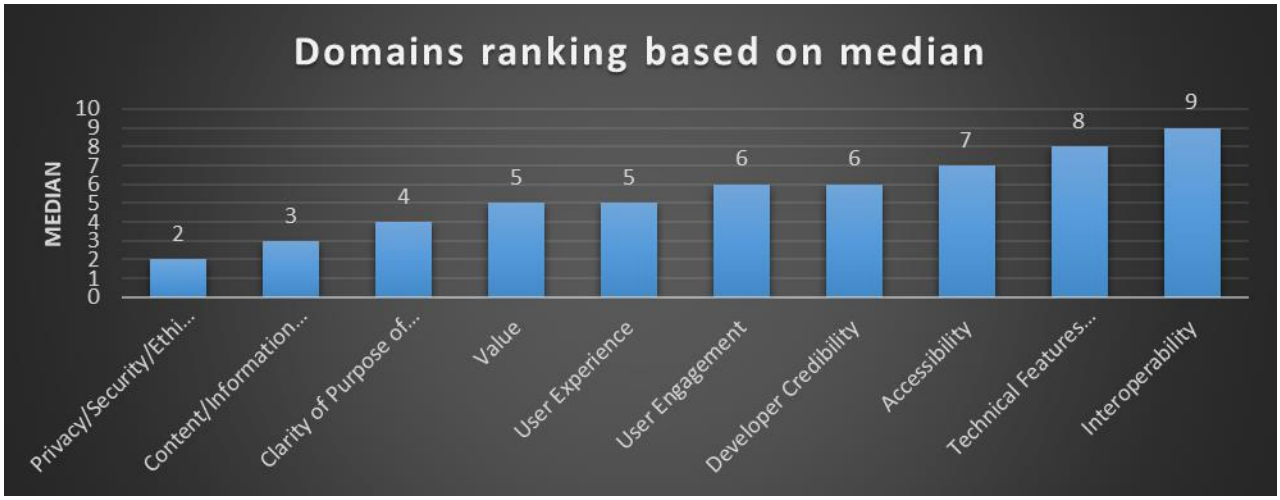
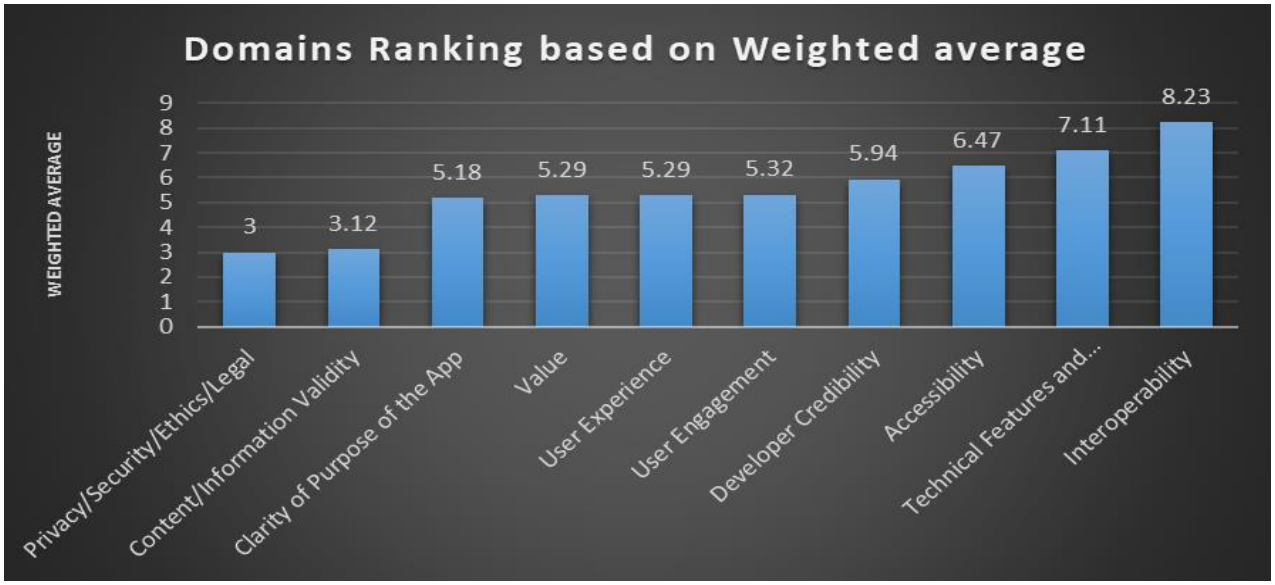


Figure 3: The order of the domains based on the weighted mean of ranking



Many participants argued that *Privacy/Security/Ethical/Legal* should be the most important domain from an organisation perspective. One participant reported that it should be the organisation's responsibility or the developer to screen this domain on behalf of end-users (hereafter, end-users refer to the health app end-users). Participants noted that end-users do generally not read the terms and conditions due to the length and detailed information embedded in these documents. One participant also noted the need to consider the individuals' capacity or the ability of the end-user to understand the privacy-related documents within an app (I.e., end-users of mental health apps).

Value domain was seen as highly important during the focus group discussion from both the end-user and organisational perspectives. The immediate outcome or benefit of the apps is the end-user's priority, thereby bypassing the security/privacy concerns associated with the health apps. The aggregated individual health benefits resulting from health apps were also valuable from the organisation perspective, noting that health gains would generate a more significant savings for a company. Further, some participants were undecided on the relative hierarchy of domains, and some reported that some domains are equally important.

"Ranking is hard as there are some criteria which should be weighted equally such as Value and Developer Credibility."

3. Practical approaches to evaluating domains based on the focus group discussion

3.1 The real definition and importance of Accessibility

The participants mentioned that *Accessibility* should not be only about language but also about how an app is accessible to vulnerable populations and people with difficulties.

"Accessible doesn't mean language only, it means for vulnerable population and people with difficulty. How to help them to access the app. Language is just something apart of it."
(Participant 13)

3.2 Valid source of information

The participants noted that health apps should be anchored on evidence-based information. Most of the participants noted that the gold standard for evaluating the validity of information is by looking at whether the health app has undergone a randomised control trial (RCT). One participant also noted that the level of evidence underpinning the information in an app could

be ranked based on RCT as the highest level of evidence. Also, the contents of health apps must be based on the clinical guidelines relevant to the health app's area (e.g., mental health and physical activity). Contents could also be supported by the guidelines of the government and experts' involvement.

3.2.1 Future possibility

A participant mentioned that carrying out validation studies on the clinical effectiveness of using health apps is a future research possibility.

3.3 Developer's Credibility

3.3.1 Who is the main driver of the health apps

Participants stressed the importance of the people or the organisation driving the content of the health apps (I.e., the organisation that influences what information to be included in an app or who commissioned an app). Some participants reported looking for the funders and organisations behind the health apps (either public or public). One participant reported that it is difficult to distinguish the developer from the funding source. The discussion on *Developer Credibility* highlighted the importance of clearly defining the term 'developer' under this domain in an evaluation framework because people define 'app developer' differently based on their disciplines.

3.3.2 Reputation and health app affiliation

The reputation of the maker of the apps was noted as a criterion to evaluate the *Developer Credibility* domain. One participant recommended reviewing the history of the app maker or reviewing the organization that promoted the app. The participants noted that an app developer or organisation with a good track of producing or recommending quality apps would also likely recommend or produce new quality apps in the market. Also, participants assumed that credible organisations, such as a trusted research organisation could prove that an app aligns with evidence-based information and best clinical guidelines and practices. Participants also reported that the app developer's credibility could be reviewed through whether the app makers are affiliated with a known organisation specialising in that field or developed by bodies of experts. Some participants reported that apps should be peer-reviewed or confirmed by a health professional to ensure it is clinically valid.

3.3.3 Divergent views in evaluating Developer's Credibility

Some participants argued that the credibility of the maker of the app should not be the basis of judging a quality app. They indicated that credibility for the same organisation varies according to various aspects. For example, a credible organisation for technology may not be credible for health app content. One participant even reported that two well-known and credible tech organisations failed to create "good health apps". One participant also stated that an app developer could "buy credibility".

"I don't think just because it's a credible developer that it's necessarily a good app or has been evaluated rigorously. So, to me, strong evidence that It's been evaluated using as higher quality evaluation framework is number one. So, I put Value down as one and, in fact, developer reputation quite further down." (Participant 4)

Further, some participants commented that they would initially refer to the developer's name or the organisation's reputation as one of the key aspects of identifying quality apps in the market. It was reported that end-users feel more secure with a trustworthy company. One

participant also alluded to the importance of the reputation of the app developer in an unregulated environment.

3.3.4 Transparency

The participants highly valued the transparency of app makers in terms of publishing information in an online space. One participant also reported that evaluators could review the organisation's background through a website search. It was brought into a discussion that some major software companies (Microsoft and Google) failed in producing quality health apps and did not have a spotless track record when it came to producing health apps. Hence, the developer's credibility needed to be assessed explicitly regarding the specific experts who produced the app's content.

3.4 User Engagement

3.4.1 Gamification and User Engagement

Some participants mentioned that gamification is one of the elements of an app that makes an app stand out from the rest in the market. Some participants said gamification makes them continuously use or engage with an app. There was also a different view about the effect of gamification functionality as it could potentially create addictive behaviour causing further harm to health.

3.4.2 User Engagement and content personalisation

The participants discussed how the audience for the content and its relevancy are important for an app's *User Engagement*. One participant mentioned that end-users tend to appreciate apps that offer personalised content.

3.4.3 Measures of User Engagement

One participant indicated that one way of measuring the real use of app and engagement is conducting an RCT or looking at previous studies. One participant added that the number of downloads does not accurately measure engagement as some end-users would download the health app but not use it. The participant stated that other measures would be the number of people who downloaded and logged in or registered, how often end –users' access an app, and how long they stay on the app. Another participant also mentioned that the frequency of activities within the lifecycle of an app is important such as how many people completed challenges/goals set up in a health app. However, they did not further elaborate on how to approach these measures.

One participant stated that checking whether the app has customer reviews or opinions verbatim is a good indicator for evaluating *User Engagement*. However, the evaluator needs to check the authenticity of the verbatim reviews as some are made-up. The participant further clarified that verbatim feedback is usually a sentence structure that provides all the information on how the end-user thought or felt about the health apps.

3.5 Privacy/Security/Ethics/Legal

3.5.1. Complexities in evaluating Privacy/Security/Ethics/Legal

The focus group discussion showed some complexities in evaluating this domain such as the capacity of end-users to review privacy statements and who has the primary responsibility of protecting *Privacy/Security/Ethics/Legal* aspects of health apps. A participant mentioned that sometimes privacy statements can be too lengthy and could be even hard to understand. Another participant stated that most end-users do not spend time in reading privacy statements available on health apps. In addition, all participants agreed that *Privacy/Security/Ethics/Legal*

responsibility lies on the organisation/app driver rather than the end-user. Another participant mentioned that ethics is the most important aspect from an organisational perspective. They also stated that it would be better if the organisations complied with the national standards and regulations. They also indicated that end-users expect all *Privacy/Security/Ethics/Legal* matters to be addressed by the app developer, requiring no further action.

3.5.2. Transparency

One participant indicated that developers should provide a complete and precise description of what data is being collected, the process of removing data, and where data is stored (I.e., locally, or overseas).

3.5.3. An area for isolated regulation

Participants noted that Australia should have specific *Privacy/Security/Ethics/Legal* regulations as some apps receive sensitive end-user information. Participants discussed that this domain should be regulated, and specific standards should be set for developers.

"That's probably an area where there could be some sort of regulatory Framework because You could have like a rate a star system or like a it is saying Australia this app meets the standards set out in and whatever. So that's it's a standardized sort of requirement across the industry in Australia." (Participant 6)

3.6 User Experience

Most of the participants agreed that *User Experience* is hard to measure by another person on behalf of the end-user. However, the *User Experience* could be explored or investigated using users' opinions. Another participant stated that end-user recommendations of apps to others are a good indicator of *User Experience*. When they were asked about the best ways to measure *User Experience*, a participant indicated W3C (World Wide Web Consortium) best practices; however, he/she did not elaborate on how it is applicable in the health apps area.

3.7 Value

3.7.1 The concept of Value

Many participants reported that the *Value* domain should be clearly defined as it is different depending on the perspective of the individual end-user, organisation or authorities. From an individual level, value means the perceived benefits of the apps. Some participants reported that value could be related to an app's clinical benefits, such as better management of chronic conditions. One participant reported that the *Value* domain could not be measured by its monetary value as some apps can be used for free but still can meet individual expectations/ purpose of the app. One participant also noted that some end-users perceived purchased apps as more valuable than free apps. Moreover, most participants reported that it is difficult to measure the *Value* domain and needs some validation study. As a proxy to evaluate this domain, some suggested evaluating the app's usability or whether it addresses the end-user's needs.

4. Relationships between domains

4.1 The link between User Experience and User Engagement to Value

One participant indicated that *User Experience* is value driven. Individuals assess value based on their experience or satisfaction with the app and whether the health apps align with the end-user expectation and intended purpose.

A participant reported that gamification directly relates to the *Value* domain, making the app more realistic for the end-user. Another participant also supported the same idea, stating that gamification and customisation are all about making an app more usable, helpful, and desirable.

4.2 The link of *Accessibility* to *User Experience* and *Value*

When participants were asked how best *User Experience* can be measured, one participant indicated a relationship between *Accessibility* and *User Experience*. Another participant also agreed, indicating that *Accessibility* and usefulness can be indicators to measure *User Experience*. *Accessibility* was also mentioned as an indicator to assess *Value*.

4.3 The link between the clarity of intended purpose from client organisational perspective and *Value*

Participants also emphasised the importance of clarity of the intended app when they were asked about the most important factor to look at in a particular app. For example, an organisation may be keen on introducing an app that attracts customers rather than looking at how valuable that is for the customer. Therefore, determining a curated set of apps could also depend on the scope or need of the client organisation (I.e., health insurance company).

4.4 The link of all the domains to Value

One participant mentioned that assessing other domains for an app would ultimately mean assessing the *Value* of the app.

Apps evaluation

COVID-19 apps evaluation during Phase I

COVID-19 apps were evaluated in the early stages of the project using the preliminary framework Version 1. A search was conducted in Apple Store, Google Play Store, Google search engine and Similarweb from 24 March and repeated weekly to 13 May 2020 to ensure that all likely COVID-19 related native mobile health apps were identified for evaluation. At the final testing of COVID-19 apps on 13 May 2020, 135 COVID-19 native mobile apps[1] were screened through four steps (after duplicates were removed) from an initial 185 apps. A total of 75 apps were excluded in Step One as they were not in English or could only be used under the close supervision of clinician(s). The remaining 60 apps were screened, with 44 apps being excluded, mainly because the target user was not the general Australian population or because the app could not be accessed or used in the Australian context due to different clinical protocols or guidelines used in the app. The remaining 16 apps were examined in Step Four against domains of the D-HEAL Framework Version 1.1 (see Appendix 1.1) and were scored according to the rating matrix. Based on our evaluation as of 20 May 2020, we recommended '*Coronavirus Australia'* and '*Healthdirect*' together with the addition of '*My Aus COVID 19*' and '*COVID Safe'* apps as being the highest-scoring COVID-19 apps then available in Australia.

Mental health apps evaluation during Phase II

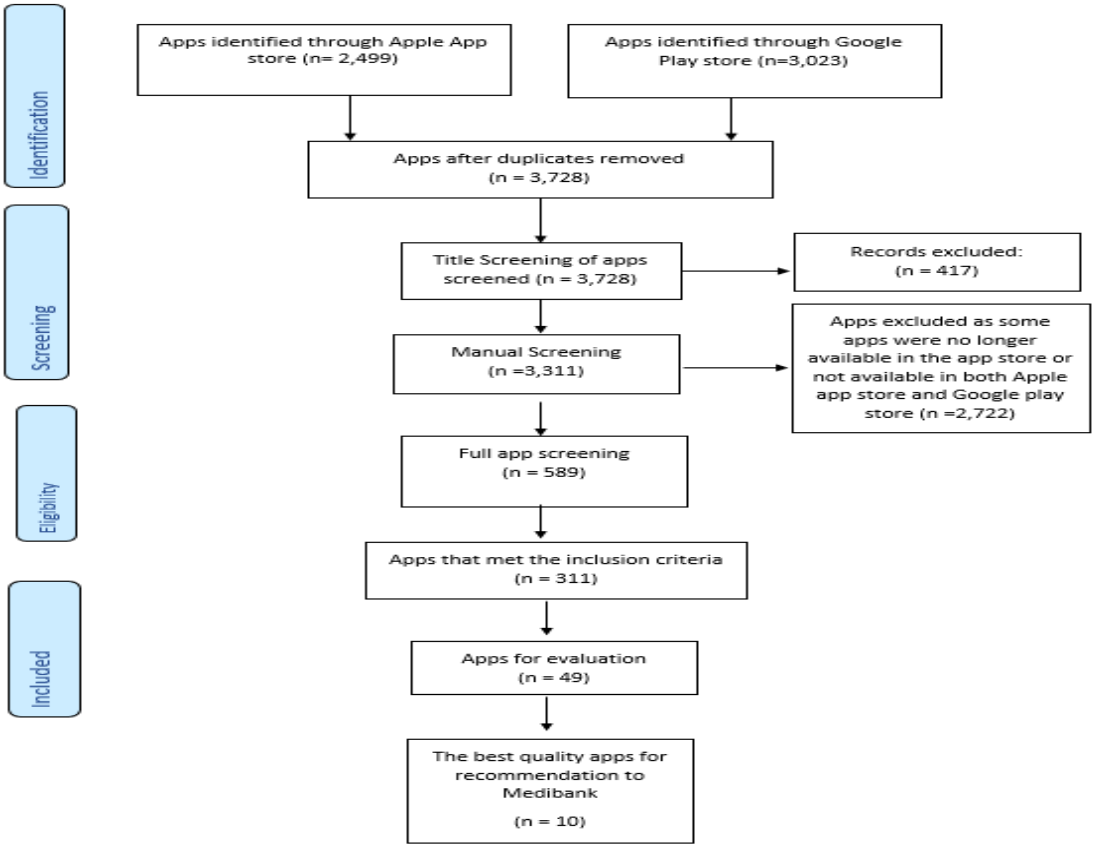
In Phase II of the project, mental health apps were evaluated using the further modified framework based on the scoping review and focus group findings. 5,522 mental health apps were found during the app search, with 3,023 apps in Google Play Store and 2,499 apps in the Apple App Store. After a manual duplicate check eliminating 1,794 apps, a total of 3,728 were reviewed for title screening. Of which, 417 were excluded based on title, and only 3,311 were reviewed for manual app title screening. 589 apps available in both Stores were included for the next screening, and 320 apps met the inclusion criteria for evaluation; however, only 49

apps were included for the app evaluation (see Appendix 3.1 and 3.2). Appendix 3.2 shows the 49 apps in detail with the app name and developer. The top ten apps were selected based on the average scoring of three raters for each app (see Appendix 3.3). Appendix 3.4 represents the average scores for all 49 apps. 'Clear Fear', 'Headspace' and 'Smiling Mind' were rated as the top three health apps from the evaluation.

In terms of resourcing, **the average time** duration for **screening, data collection and app evaluation per app was 3 mins (171 seconds), 5 mins (326 seconds), and 1 hour** respectively.

Figure 4 shows the flow of the process from the screening to data evaluation.

Figure 4: PRISMA flow diagram



Inter-rater reliability

Two tests were conducted to test inter-rater reliability of app rating across three raters: Fleiss's Kappa and Intraclass correlation coefficient.

Fleiss's Kappa

None of the apps were at the extremes of either poor agreement (below 0) or almost perfect agreement (0.81-1.0). Instead, **most of the apps out of 49 (n=25, 51%) showed moderate agreement (0.41-0.6),** followed by fair agreement (0.21-0.40) for 17 apps (35%). Three apps had a substantial agreement (0.61-0.8), and four apps had only slight agreement (0.00-0.20). The calculated Fleiss's Kappa for each question of the framework (three raters and all the apps subject to each question) showed that the most frequent score for questions (n=13/27) was "slight" agreement among three raters. Table 3 shows the level of agreement among three raters for each question.

Table 2: Interrater reliability for each question using Fleiss's Kappa

Agreement level among three raters for each question	# Of questions	Questions
Poor Agreement (Below 0)	1	Q17
Slight Agreement (0.00-0.20)	13	Q1, Q2, Q3, Q6, Q7, Q8, Q9, Q10, Q13, Q15, Q19, Q24, Q26
Fair agreement (0.21-0.40)	7	Q4, Q5, Q11, Q12, Q18, Q21, Q27
Moderate agreement (0.41-0.60)	4	Q16, Q20, Q22, Q25
Substantial agreement (0.61-0.8)	1	Q14
Almost perfect (0.81-1.0)	1	Q23

Intraclass correlation coefficient (ICC - Cronbach Alpha)

The pooled analysis that accounted for all apps (n=49) showed a high consistency of observation amongst raters and apps. The ICC average measure for all observations was .855, 95% CI (.819-.886) (Alpha was within range .75-.9, indicating good consistency in agreement). Observation of single items indicated a high level of consistency (overall Alpha score > .8 if a single item was deleted). Opportunity for improvement may exist by removing questions 6,16,17,18, and 20. However, this needs to be confirmed by analysis per apps and domains approaches. See Appendix 4 for in-detail information for the ICC subgroup analysis. Among 49 apps rated by three raters using 27 questions for ten domains, ICC analysis for individual apps indicated a statistic of up to 61% of rated apps with a high level of consistency in rating. The graph below showed the distribution of the 30 apps, which displayed a "high" reliability score with a 95% confidence interval (Figure 5). Analysis stratified by questions showed the distribution of poorly rated questions in the poorest rated apps (Figure 6). In consideration for items within each domain, all items within *Privacy/Security/Ethics/Legal* appeared to have unreliable items showing a poor agreement for this domain. In contrast, the opposite was observed in *Content/Information Validity*. *Privacy/Security/Ethics/Legal* domain consisted of questions (9-12) had the poorer average ICC result in poorly rated apps (n=7/10) and good rated apps, suggesting improvement is required for this domain. Due to the equivocal result for *Privacy/Security/Ethics/Legal* in this exercise for this Framework Version 2, this domain as an independent domain is more viable option due to the acceptable/good consistency in other domains.

Compared to Fleiss's Kappa analysis, the overall result was consistent with some absolute agreement detected by the two app and item consistency rating tests. However, some apps were in disagreement and had a negative ICC score, suggesting the sampling error that attributed to a small sample size (i.e., number of available raters) as noted above.

Figure 5: Good reliability rated app with 95% Confident Interval

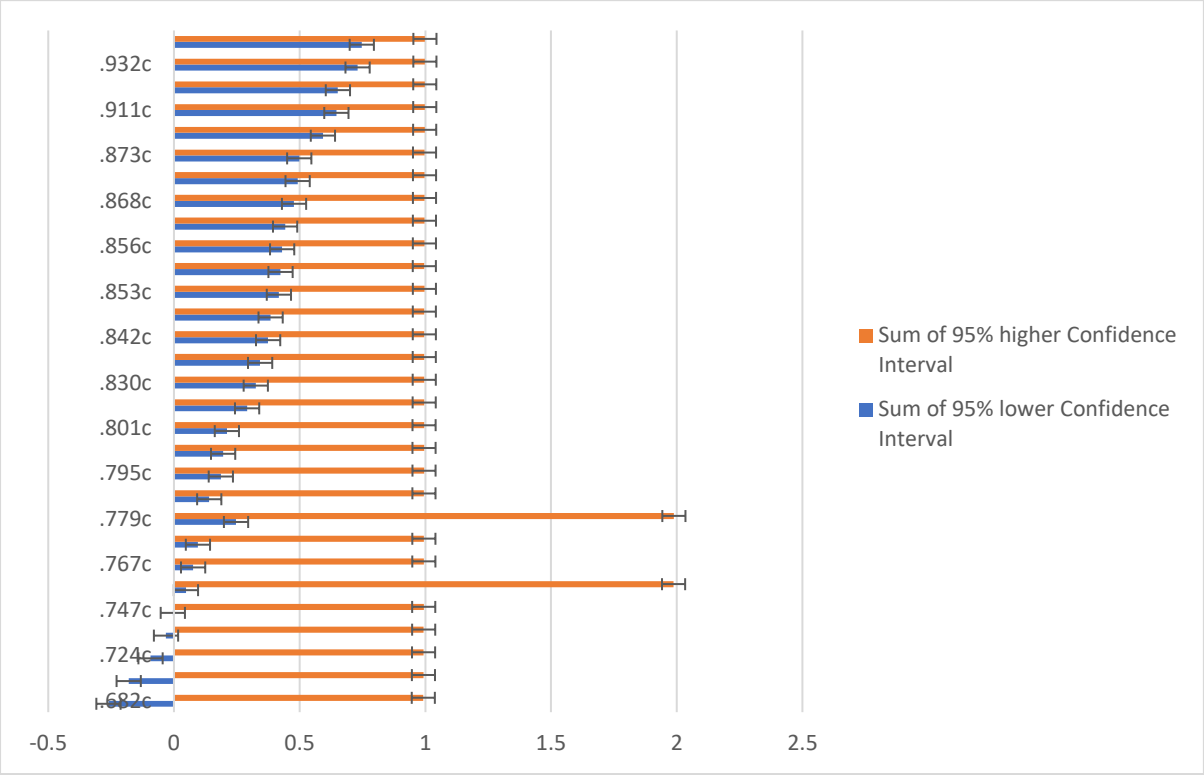
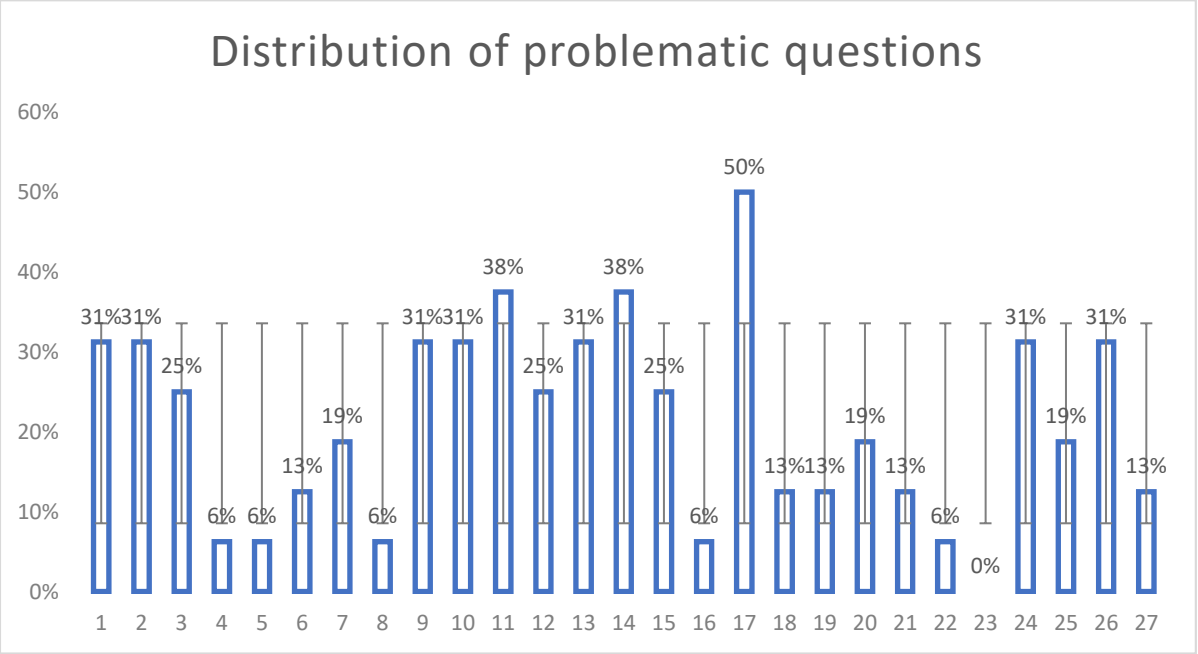


Figure 6: Distribution of problematic questions according to poorly rated apps



Survey data Analysis

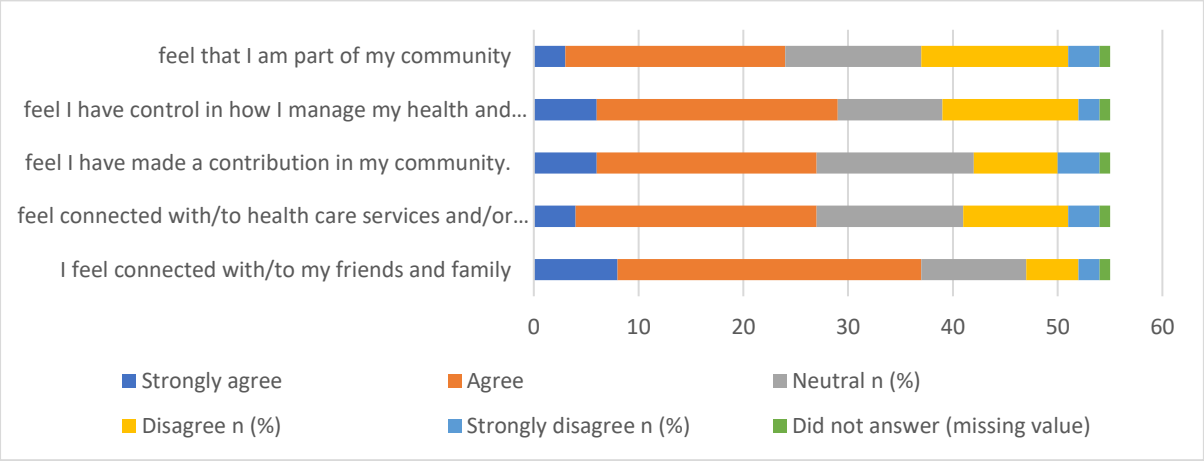
Seventy-eight people responded to the baseline survey (S1), but 23 did not complete S1 beyond providing consent. The majority fully completed the survey (n=42, 53.8%), while 13 completed it partially (16.7%), who were also included in the analysis.

65.4% of the total S1 respondents dropped out before completing the follow-up survey 2 (S2). Twenty-seven people responded to S2, and one response was incomplete. Drop out was reduced between S2 and S3. Of those who responded to S2, 85% (n=23) also completed survey 3 (S3).

Appendix 5 indicates the demographic details of the responders. The majority who responded to the baseline survey were female (n=48, 87.3%), which is striking. Of those who completed S2 (n=27), only one was male, and the rest (~96% in both) were females. Across surveys S1, S2 and S3, there was not much variation in participants' education, and employment distribution; most were in age groups 26-35 and 46-55 years old, more educated, and worked full-time. More participants had previously used health apps than those who had not previously used any.

In terms of the health status of respondents, generally, this cohort reports good health status and are very motivated to work on mental health wellbeing, which they thought is very important. Regarding their current circumstances and state of health and wellbeing, the largest group did not indicate any current distress and seemed contented with their current health and wellbeing. (Figure 7)

Figure 7: Level of current circumstances and state of health and wellbeing



61.8% of the respondents had used a mental health or wellbeing app in the past. The majority reported that they had not paid anything for the health apps they had used in the past. Among those who had used such mental health and wellbeing apps previously, the most frequently used app had been 'Smiling Mind' (n=15) followed by 'Headspace' (n=6) – both of which were, of course, also two of the three apps recommended by Medibank as part of this study. Based on weighted average rankings, *Content/Information Validity* was seen as the most important domain by all the participants (Figure 8) and those who had previously used health apps (Figure 9), whereas *Interoperability* was ranked as the least important domain. This finding was consistent with the focus group findings. However, the rankings by those who had not previously used health apps differed significantly, with non-users ranking *Technical*

Features and Support as the most important and *Developer Credibility* as the least important domains (Figure 10).

Figure 8: Weighted average of Ranking of the Domains by all the 55 respondents

(34 previous app users, 11 non-app users, and 10 participants with missing information about previous app use)

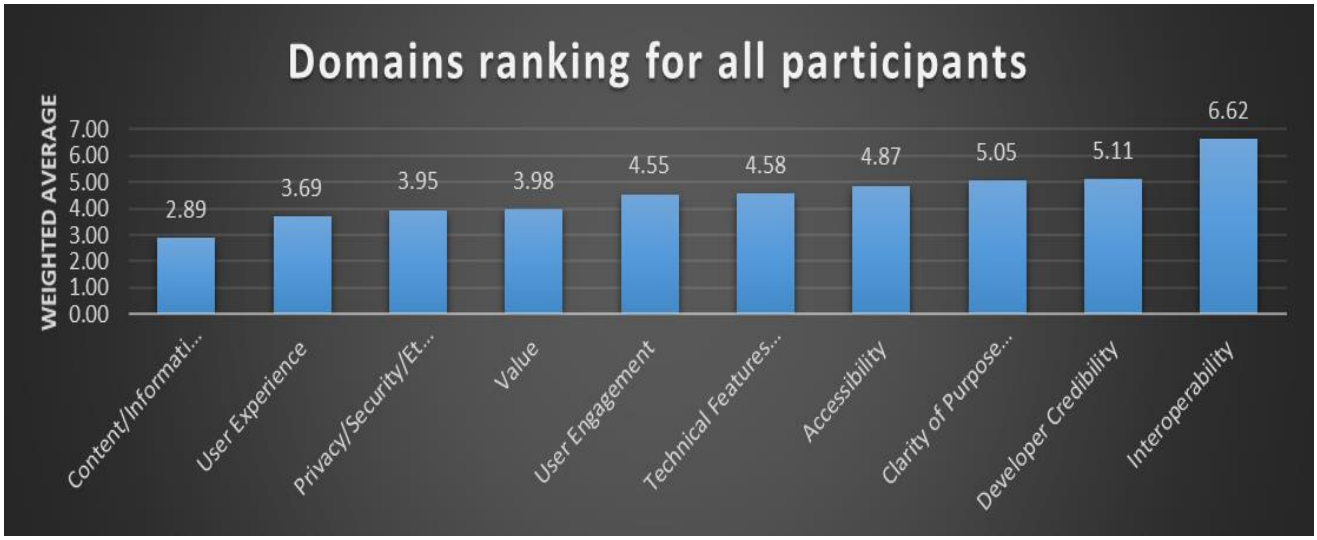


Figure 9: Weighted average of Ranking of the Domains by the respondents who had used a health app before (Previous app users)

(34 respondents)

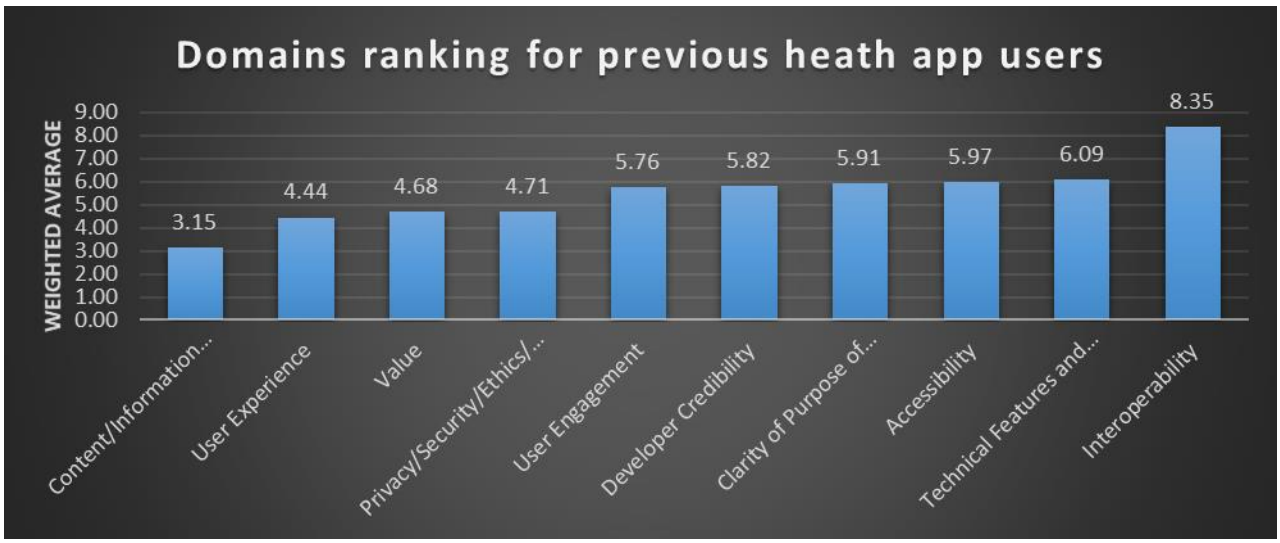
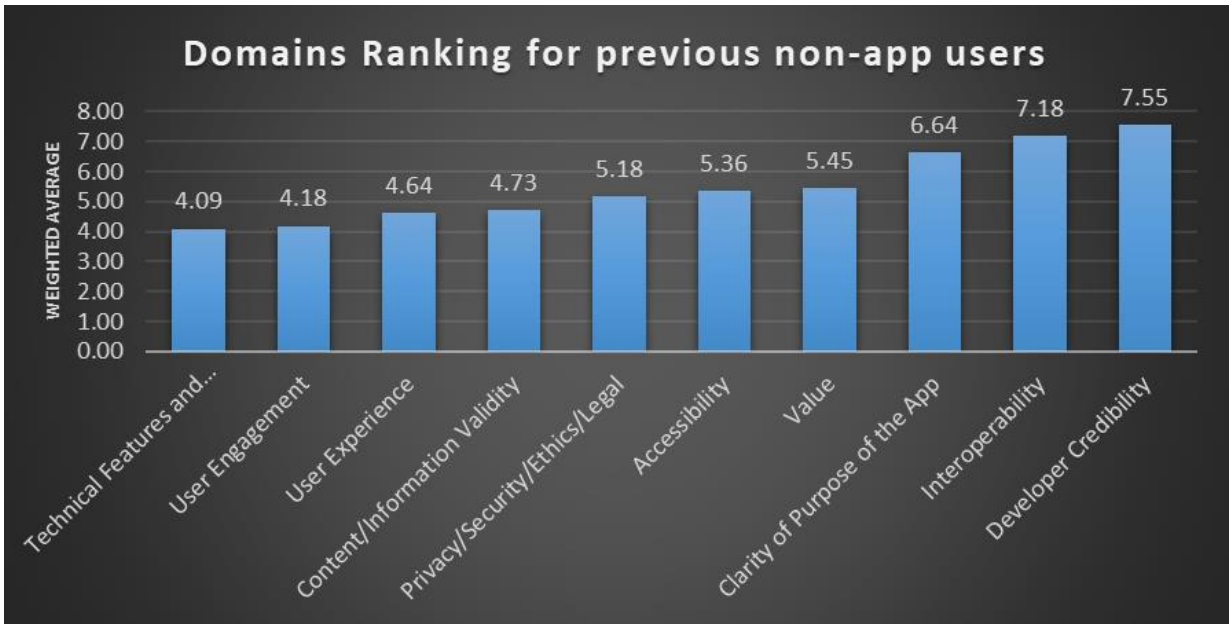


Figure 10: Weighted average of Ranking of the Domains by the respondents who had NOT used a health app before (Previous NON-app users)

(11 respondents)



There was no significant correlation between income and willingness to pay at baseline or the end of the follow-up survey 3. However, there was a significant correlation between frequency of use and willingness to pay by the end of Survey 3 (p-value = 0.0444). A linear regression established that the frequency of use could statistically significantly predict the willingness to pay for a health app, p=0.0444. The frequency of use accounted for 13.96% of the variability in the willingness to pay.

Fifteen respondents (57.6%) in S2 were unwilling to pay anything for health apps (AUD$0). The mean willingness to pay among those who answered S2 was AUS$8.35 for 26 respondents, and it was AUS$5.04 for the 23 respondents who completed S3. In S3, 12 respondents (52.1%) were unwilling to pay for health apps (AUD$0). The mean willingness to pay was reduced between S2 and S3 but not statistically significant. A similar pattern in willingness to pay can be seen between Baseline and S3. The change is not statistically significant from Baseline to S3. Twenty-two (40%) respondents at baseline were not willing to pay anything, and the mean willingness to pay at baseline was AUS$5.71 per month among 42 respondents who responded and AUD$5.04 at S3 for 23 respondents.

Overall, the mean ratings on the Likert scale for health app usability questions increased from S2 to S3. The change in the scale increased positively for 58% (11/19) of questions, which indicated that usability and satisfaction with the apps increased over time. All three apps had the same pattern of positive change (i.e., ‘Smiling Mind’ 15/19, ‘Headspace’ 14/19, ‘Clear Fear’ 11/19). 37% of the questions (7/19) had negative mean change, while 10.5% (2/19) had no change over time. Respondents reported that their frequency of app use reduced over time for all three apps. All the apps were rated as easy to use and easy for the user to learn how to use the app, and navigation was convenient moving between screens. The Likert scale rating increased over time for these features. App users were more satisfied with ‘Smiling Mind’ than ‘Headspace’, and our framework also rated these apps with a high score for user

satisfaction related domains such as *User Engagement and User Experience*. 'Clear Fear' was the least satisfied app, but the scoring from framework indicated the opposite, where this app scored high for satisfaction related domains. It should be noted that only two app users used and completed the surveys on 'Clear Fear', which was the reason for opposite results for this app. Based on the open-ended questions, participants' expectations from using a health app were generally managing their mental health status. At individual levels, they were more specific, and some of the expectations were managing stress, clearing fear, better understanding, and improving their mental health. Interestingly, Headspace and 'Clear Fear' users reported a reduction in their app rating over time as beneficial for their health and well-being, but the change is not significant. Their rating of the apps as improving access to healthcare services increased for all the apps between S2 and S3.

Our prior evaluation of 49 mental health and wellbeing apps using our app evaluation framework had resulted in 'Clear Fear' as the first ranked app, followed by 'Headspace' and 'Smiling Mind', but only two participants selected 'Clear Fear' app for use and evaluation in the survey. This low number of users for 'Clear Fear' could be because this app is mainly recommended for young people between the ages of 11-19 years, and our survey respondents in that age group are limited. This could also be because of the participants' awareness of the 'Headspace' and 'Smiling Mind' apps, as Medibank promoted these two apps. Feedback for all three apps was generally positive. Nine out of ten respondents who used 'Smiling mind' recommended this app to other members. Similarly, eight out of 11 respondents recommended 'Headspace' to other members.

Reasons for those who did not recommend Headspace was that the app was expensive (i.e., one participant mentioned that "I felt there were too many things that were locked unless you'd paid for the app, and while I've enjoyed it when I've used it, I don't think I'd get enough benefit from it to pay for it.") One respondent did not recommend 'Smiling mind', indicating that it does not work, which may be related to a technical issue or similar situation. A respondent who did not recommend 'Clear Fear' in S2 changed their recommendation in S3, indicating that the app is recommended. The respondent stated that it is because "It may help someone or give them a better idea of what they don't want and to look elsewhere for something easier to navigate, track feelings & interactions with the app. I liked the bubbles." The respondent who changed their app from 'Smiling mind' to 'Headspace' from S2 to S3 recommended the app in each stage.

Reasons for recommending 'Smiling Mind' app from respondents' viewpoint included: Useful, user-friendly, easy to use, easy to understand, free components, a wide variety of meditations, record moods that allow comparison, downloadable meditations, a great range of resources and activities, Australian accent guided meditation, valuable services at an additional cost, and cheaper in monthly payments. Reasons for recommending 'Headspace' app from respondents' viewpoint included: Easy to use, useful, availability of practical assistance, appropriate meditation, a great range of meditations, free videos, practical, and enjoyable. Reasons for recommending 'Clear Fear' app from respondents' viewpoint included: clear avenues for seeking support and a useful and accessible interface.

Framework revisions from the Workshop

On 09 and 16 February 2022, five researchers (MH, PC, SWAD, MRSA, DN) met to improve the framework according to the user evaluators' feedback, inter-rater reliability (IRR) output, focus interview and user survey data analysis. Figure 11 shows the general structure of the question items in the framework, and Table 3 shows the summary of the items updated. Appendix 1.2 presents framework Version 2 and detailed changes to develop Version 3.

The majority of the questions were simplified or reworded to improve the agreement. Some subjective words such as "useful", "limited", and "quality" were removed. The researchers also included notes or more descriptions to further guide the framework's users. Five questions were simplified to add more clarity. The criteria within the questions were updated or reworded for ten questions. Answer options were updated or reworded where relevant (Table 9). To avoid scoring issues, the scoring values assigned to answer options were updated for three questions. Based on one of the vital feedback items, the answer options for questions were revised and updated to be able to score.

The name of four sub-domains was updated to reflect the correct representation of the questions asked. Three sub-domains were deleted (perceived value, ease of use, and customer services) as these were noted having only slight agreement in the IRR analysis. Moreover, these items were deemed highly subjective. Removing perceived value sub-domain indicated that *Value* domain needs to be excluded.

Four key items need further investigation in future studies. Firstly, the IRR result for the Privacy*/Security/Ethics/Legal* domain was problematic. The importance of the *Privacy/Security/Ethics/Legal* domain is one of the key findings from the literature review and the focus group interview. However, we found that it is the hardest and the weakest area of the framework to evaluate. Therefore, we assessed this domain separately without incorporating it into the overall scoring but still keeping it as a domain in the framework. Another finding was that although the Vulnerable population sub-domain under the *Accessibility* domain was considered important by the focus interview participants, assessing this sub-domain was complicated. This question asked whether there are any support features in an app to support those with limited digital health literacy and visual or hearing impairment. There was only a slight agreement among raters on this question. The raters were often confused about differentiating the visual and hearing features available in the "apps themselves" and the device. Online functionality (Q18) was another item for further improvement as some apps do or do not allow real-time communication (e.g., peer support apps) and thus needs to be accounted. Finally, star ratings as a proxy for *User Experience* have an "almost perfect" agreement. However, users of the framework noted a potential "ceiling effect". One recommendation was to add another question asking for the number of people who rated the app or the number of app downloads. In the final step of Phase II, the framework was finalised with nine domains and 24 questions.

The screening section needs to be updated accordingly:

1. Filter items such as broken developer weblink or app website need to be included.

2. Consideration to review the availability of data privacy features in the Apple store
3. Cessation of app evaluation if T&C was not available or if T&C is not in the English language
4. Usage of generic email during app registration (if needed)
5. Consider screening out USA apps.

Some evaluator feedback was not incorporated as some comments were specific to the apps evaluated. Proper training can address some comments, and IRR results are already assessed as fair to a moderate agreement.

Figure 11: The structure of question items in the framework

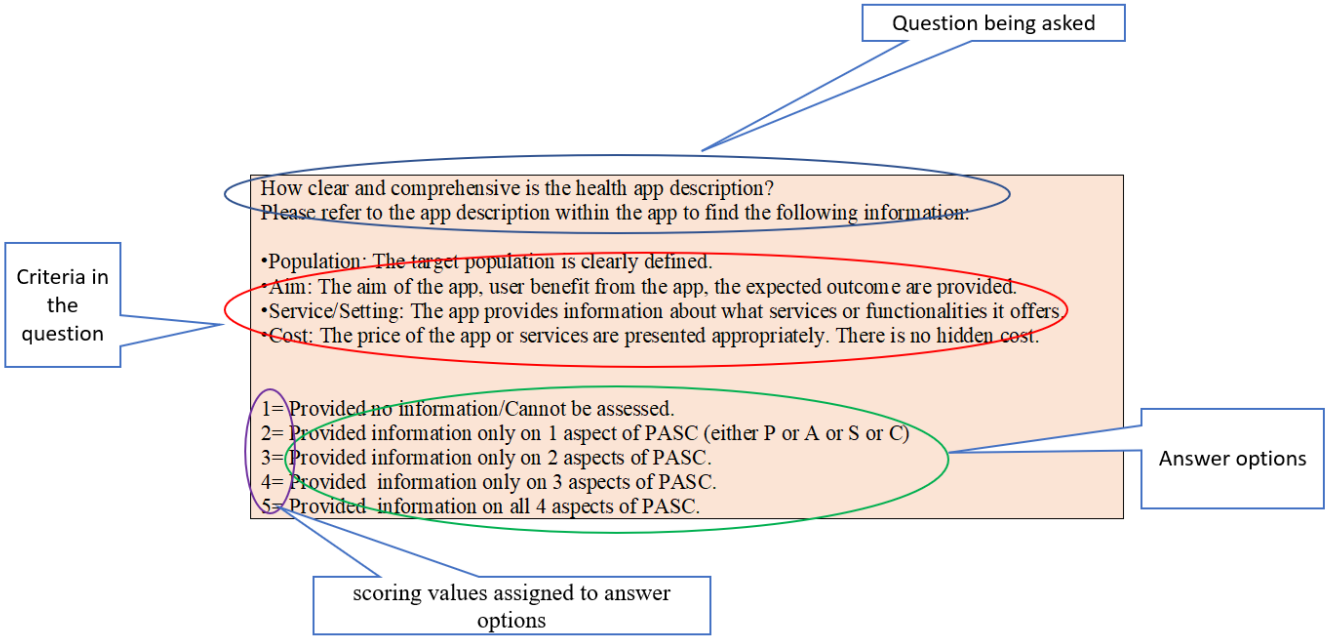


Table 3: Items in the framework that were updated during the final revision

Updates	**Question Number**
Reworded the question being asked	Q3, Q7, Q8, Q17, Q20
Reworded/updated criteria in the question	Q2, Q3, Q5, Q6, Q7, Q17, Q19, Q22, Q25, Q27
Reworded/updated answer options	Q1, Q2, Q3, Q4, Q5, Q7, Q8, Q9-Q12, Q13, Q14, Q17, Q18, Q19, Q20, Q22, Q23, Q27
Updated scoring *Value*s assigned to answer options	Q2, Q19, Q27
Sub-domain name	Q7, Q8, Q17, Q22
Deleted	Q24: Perceived *Value* (*Value* domain) Q26: Ease of Use Q15 Customer Services
Need further investigations in future studies	Q9-Q12:Privacy security ethics legal domain

	Q18: Online functionality Q19: Vulnerable population Q23: User ratings

DISCUSSION

The aims of the project included developing a health app evaluation framework to support consumers to choose high-quality health apps and evaluating the usability of this framework when applied to apps covering general health issues frequently encountered in the general population.

This project included two phases with multiple research components to address the project's research questions. The first phase included conducting a scoping review, developing a framework, and conducting focus group interviews with health experts. The second phase involved assessing mental health apps using the draft framework, testing inter-rater reliability, estimating the time cost for using the framework by organisations, assessing the reliability of framework rating compared to app users' perceptions collected using validated survey instruments.

One journal article has been produced out of this project, which is the scoping review. The scoping review findings indicated that there are overlaps between some components of domains, suggesting there is some flexibility within frameworks and that categorisation of domains is not yet a standardised process. There is a significant amount of literature available in this space, and the methodologies for scoring domains and the way domains have been utilised widely varied across already published frameworks. Four hundred and thirty assessment criteria (questions/items) were compiled across 97 identified studies. The selection of questions from the list for an app evaluation framework should be carefully conducted based on criteria including, but not limited to, the structure, the depth and the expected outcome from the question, and the subjectivity or objectivity, because individual perceptions on the quality of the questions can vary from one person to another. The most frequently reported scaling method was a point system: the 5-point Likert scale. The most popular approach for scoring was the calculation of mean/average scores. Some studies used different weightings of domains, while some frameworks did not use a scoring method but used either a pyramid style model or a descriptive evaluation. Based on the most frequently used methods, we adopted a 5-point scale but calculated the total score.

A framework was first developed alongside the scoping review based on the synthesized domains and items of the recently published literature found from preliminary searches. This framework was further refined along the way, producing Version 2 and the Final Version 3. The first framework included seven domains, then developed into ten domains as the next step with 27 criteria (Version 2). In the final step of Phase II, the framework was finalised with nine domains and 24 questions (Version 3).

When analysing the evaluators' ratings, Fleiss's Kappa statistical analysis showed that most of the apps out of 49 (n=25, 51%) showed moderate agreement (0.41-0.6), followed by fair agreement (0.21-0.40). Compared to Fleiss's Kappa analysis, the overall result of ICC was consistent with some absolute agreement detected by the app and item consistency ratings. These results suggest more consistency across different raters of our framework. However, some apps were in disagreement and had a negative ICC score, suggesting a sampling error

that could attribute to a small sample size. In terms of resourcing, the average time duration for screening, data collection and app evaluation per app was 3 mins (171 seconds), 5 mins (326 seconds), and 1 hour respectively.

There were key findings for some domains due to various stages of the project. Focus group interviews found that the *Privacy/Security/Ethics/Legal* domain and *Content/Information Validity* domains should be prioritised as the most important, while *Technical Features and Support* and *Interoperability* were viewed as the least important. This is relatively consistent with the survey findings where *Content/Information Validity* was ranked as the most important domain, and *Privacy/Security/Ethics/Legal* domain was also among the top ranking. Moreover, *Technical Features and Support* and *Interoperability* were rated as the least important domains in the survey. The importance of the *Privacy/Security/Ethics/Legal* domain is also one of the key findings from the literature review. However, it is the hardest and the weakest area of the framework to evaluate, which we found from the evaluators' feedback and inter-rater reliability testing. Therefore, further investigation is needed in future studies for this domain. Our report highlights the importance of the app stores and app developers to abide by requirements related to consumers' privacy concerns.

There were consistent results on *User Experience* across focus group interviews and survey data, which indicated that this domain is harder to evaluate by an app evaluator within a limited time, given this requires a considerable amount of time to experience the app.

Although the focus groups interview results indicated the important need for exploratory research to investigate the perceived value to end-users and to provide a clearer picture on how to assess the *Value* domain, our consumer survey allowed us to capture the perceived value of end-user in terms of how useful and effective an app for them through their expectations of the app and their recommendation. However, incorporating perceived value in the evaluation framework is problematic as it is hard to assess by an evaluator. This is also the case for both easy to use and customer service sub-domains.

Accessibility was identified as a new domain deemed crucial for assessment due to ethical and equity considerations. The benefit of *Accessibility* as a measurable domain arose in consideration of the framework on COVID-19 apps, where public criticism of the apps was their poor support for CALD communities. In addition, the importance of *Accessibility* emerged in the focus group discussions. Our study finds that in the mental health and wellness area, it is imperative for apps to support multiple languages and have simple user interfaces that have been ethnographically tested to ensure usability in a mixture of communities and people with disabilities. Such aspects of fairness and equity are infrequently considered in the evaluation frameworks, and nor are they a feature of many major apps on the marketplace. Our report highlights the need for the requirement by the app stores to require clear evidence of *Accessibility* support for apps specific to the health domain.

CONCLUSION AND RECOMMENDATIONS

To conclude, the framework initially synthesised ten health apps assessment domains. These were primarily based on the evidence from our scoping review: Clarity of purpose of app, *Developer Credibility*, *Content/Information Validity*, *User Experience*, *User Engagement*, Interoperability, *Value*s, *Technical Features and Support*, *Privacy/Security/Ethics/Legal*, and *Accessibility*. The framework adopted the most commonly used 5-point Likert scale as a scaling mechanism and the total score for scoring. The average score of the total was used when multiple raters evaluated the same health app. The final draft of the framework consists of nine domains with 24 questions after incorporating all the findings from each step of phases.

There are overlaps between some components of domains, suggesting there is some flexibility within frameworks and that categorisation of domains is not yet a standardised process. Therefore, we reviewed the ten domains based on scoping review, the focus group study, and the evaluation process.

Even though *Value* domain was deemed important based on the scoping review and focus group study, this was found to be a highly subjective domain to assess during the evaluation and therefore removed from the final framework. Removing this domain could justify using focus group findings that reported assessing all the other domains would mean that *Value* of the app has been assessed.

Similarly, the Privacy/security/ethical/legal domain was more complex to assess. Therefore, the scoring of this domain was not incorporated into the total scoring for health apps. Moreover, locating information to assess this domain needs thorough training. More research is needed to review *Value* and Privacy/security/ethical/legal domains.

Accessibility was identified as a new domain deemed crucial for assessment due to ethical and equity considerations. The benefit of *Accessibility* as a measurable domain arose in consideration of the framework on COVID-19 apps, where public criticism of the apps highlighted their poor support for CALD communities. In addition, the importance of *Accessibility* emerged in the focus group discussions. Our study finds that in the mental health and wellness area, it is imperative for apps to support multiple languages and have simple user interfaces that have been ethnographically tested to ensure usability in a mixture of communities and by people with disabilities. Such aspects of fairness and equity are infrequently considered in the evaluation frameworks, and nor are they a feature of many major apps in the marketplace. Our report highlights the need for the requirement by the app stores to require clear evidence of *Accessibility* support for apps specific to the health domain.

At the end of all the stages, the project produced the Deakin Health E-technologies Assessment Lab Framework for Mobile Health Evaluation (D-HEAL FMHE), comprising 24 questions under **nine** domains (Version 3). The framework was developed for organisations to evaluate health apps to promote the best health apps for their clients within the Australian context. As we tested this framework on mental health apps, we can conclude that the framework is reliable based on the consistency found across different raters during the evaluation. There are possibilities to extend this framework in future studies. The framework could be used and further tested using experimental studies in different settings. Moreover, this framework could be further developed to meet different needs, such as modifying this to suit end-users as evaluators or for different kinds of health apps that differ in functionality or purpose.

REFERENCES

1. Hensher, M., et al., *Scoping review: Development and assessment of evaluation frameworks of mobile health apps for recommendations to consumers.* Journal of the American Medical Informatics Association, 2021. **28**(6): p. 1318-1329.
2. Henson, P., et al., *Deriving a practical framework for the evaluation of health apps.* The Lancet Digital Health, 2019. **1**(2): p. e52-e54.
3. Lovett, L. *Mental health apps plentiful, but few provide clinical research*. 2019 [cited 2021 04 November]; Available from: https://www.mobihealthnews.com/content/mental-health-apps-plentiful-few-provide-clinical-research.
4. Moshi, M.R., et al., *Evaluation of mobile health applications: Is regulatory policy up to the challenge?* International journal of technology assessment in health care, 2019. **35**(4): p. 351-360.
5. Ferretti, A., E. Ronchi, and E. Vayena, *From principles to practice: benchmarking government guidance on health apps.* The Lancet Digital Health, 2019. **1**(2): p. e55-e57.
6. Department of Health therapeutic Good Administration. *Regulation of software based medical devices*. 2021 [cited 2021 15 December]; Available from: https://www.tga.gov.au/regulation-software-based-medical-devices.
7. Magrabi, F., et al., *Why is it so difficult to govern mobile apps in healthcare?* BMJ health & care informatics, 2019. **26**(1): p. e100006.
8. Australian commission on Safety and Qaulity in Health Care. *National Safety and Quality Digital Mental Health Standards*. 2019 [cited 2021 04 November]; Available from: https://www.safetyandquality.gov.au/standards/national-safety-and-quality-digital-mental-health-standards.
9. Guide, D.H. *An evidence-based approach to recommending mobile health apps and digital health solutions*. 2021 [cited 2021 04 November]; Available from: https://digitalhealthguide.com.au/DHG/welcome/digital-health-guide.html.
10. Stoyanov, S.R., et al., *Mobile app rating scale: a new tool for assessing the quality of health mobile apps.* JMIR mHealth and uHealth, 2015. **3**(1): p. e3422.
11. Council, A.H.M.A., *National Mental Health Plan, 2003-2008*. 2003: Commonwealth of Australia.
12. Nouri, R., et al., *Criteria for assessing the quality of mHealth apps: a systematic review.* Journal of the American Medical Informatics Association, 2018. **25**(8): p. 1089-1098.
13. Braun, V. and V. Clarke, *Using thematic analysis in psychology.* Qualitative research in psychology, 2006. **3**(2): p. 77-101.
14. Stata.com, *Manual-Kappa interrater agreement*. 2021.
15. Koo, T.K. and M.Y. Li, *A guideline of selecting and reporting intraclass correlation coefficients for reliability research.* Journal of chiropractic medicine, 2016. **15**(2): p. 155-163.
16. Tavakol, M. and R. Dennick, *Making sense of Cronbach's alpha.* International journal of medical education, 2011. **2**: p. 53.
17. Revilla, M.A., W.E. Saris, and J.A. Krosnick, *Choosing the number of categories in agree–disagree scales.* Sociological Methods & Research, 2014. **43**(1): p. 73-97.
18. Wisniewski, H., et al., *Understanding the quality, effectiveness and attributes of top-rated smartphone health apps.* Evidence-based mental health, 2019. **22**(1): p. 4-9.

APPENDICES

Appendix 1: Framework development

Appendix 1.1: D-HEAL Evaluation Framework Version 1

Version 1				
Version 1.0			**Version 1.1**	
Domains	**Objective appraisal (based on the quantity)**	**Subjective assessment by three evaluators**	**Domains**	**Objective appraisal (based on the quantity)**
Functionality			*Functionality*	
Pop-up notification/ alert sharing information/ Requirements to use	X		Pop-up notification/ alert sharing information/ Requirements to use	X
Provide health information	X		Provide health information	X
Data collection and data storage to support clinical decisions	X		Data collection and data storage to support clinical decisions	X
Connectivity	X		Health warnings, and goal setting	X
Health warnings and goal setting	X		Provide information of additional support service	X
User Experience, adherence & engagement			*User experience, Adherence & Engagement*	
User-friendliness		X	Design: lean and attractive	X
Engagement		X	Engagement- user interface and activities	X
Interoperability			User autonomy	X
Data sharing of personal contact information	X		information volume and presentation	X

Data sharing of location using GPS and Bluetooth	X		Accessibility- Online/offline	X
Technology Requirements			Accessibility- Individual need	X
Standalone/Use of peripheral devices	X		Accessibility - Language	X
Data Storage (in device)- MB	X		Accessibility - support strong business case- app design for a particular population	X
Compatibility across	X			
Information Validity			*Interoperability*	
Source of information	X		Data sharing of personal contact information	X
Ethics, Data privacy, Legal and Legislations			Method of Data sharing of location using GPS and Bluetooth	X
Security	X		Ability to sync data	
Privacy	X		User ability to control data	X
Age	X		*Technology Requirements*	
Laws and legislations	X		Standalone/Use of peripheral devices	X
Value was not assessed			Data Storage (in device)- MB	X
			Compatibility across platform	X
			Enabling GPS /Bluetooth	X
			Required most up to date operating system i.e.: latest IOS/additional software requirement	X
			Value	
			Tangible value (cost/outcome)	
			potential cost saving	X
			potential to improve health outcome	X
			potential out of pocket cost	X
			Intangible value (loyalty/trust)	
			potential link to external services	X
			potentially advocate equity	X
			potential improve knowledge	X
			Intrinsic value refer to the value of apps itself	X

	Extrinsic value- refer to the measure of how the apps can change your life	X
	App/Information Validity	
	Trusted organisational (International, national and Evidence based (Experimental)	X
	Evidence base practice I.e.: descriptive study	X
	Expert opinion/media release/ non-peer review research activity	X
	Internet based source- unclear validity	X
	Not mention	X
	Ethics, Data Privacy, Legal and Legislations	
	Privacy- compliance	X
	Privacy- further attention from developer on mistrust including data control	X
	Security- 2 factor authentication	X
	Security- Data code/encrypt	X
	Security- data storage	X
	Security-Data analysis	X
	Security- Data handling process	X
	Security- Attention to additional technology that impact to the measure of security	X
	Ethics- age use	X
	Ethic- consent	X
	Ethic-equity	X
	Ethic- Conflict of interest	
	Laws and legislation	X
	Laws and legislation - current framework/protocol	X
	Laws and legislation - Involvement of international/foreign policy/framework	X

	Laws and legislation - Compliant with international legal policy and/or framework	X

Appendix 1.2: D-HEAL Evaluation Framework Version 2 and 3

Domain **Sub-domain**	**#**	**Framework Version 2**	**#**	**Framework Version 3**	**Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/** **Focus Interview (FI)/ Survey data analysis (SDA)**	**Actions made to address the comments and results of the analyses**
A. *CLARITY OF PURPOSE OF THE APP*						
CLARITY	1	How clear and comprehensive is the health app description? Please refer to the app description within the app to find the following information: • Population: The target population is clearly defined. • Aim: The aim of the app, user benefit from the app, the expected outcome are provided. • Service/Setting: The app provides information about what services or functionalities it offers. • Cost: The price of the app or services are	1	How clear and comprehensive is the health app description? Please refer to the app description within the app to find the following information: •Population: The target population is clearly defined. •Aim: The aim of the app, user benefit from the app, the expected outcome are provided. •Service/Setting: The app provides information about what services or functionalities it offers. •Cost: The price of the app or services are presented	**FB:** We need to deal with apps designed for connected devices as the primary purpose of the app. **IRR:** Slight Agreement, For the items' reliability, both analyses indicated that Questions 1-3, 9-13, 15, 17, 24, and 26 are the question with poorer reliability within the domains. **FI:** *Clarity of Purpose of the App* appeared to be the most important domain	Removed "limited and detailed to address inter issues."

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		presented appropriately. There is no hidden cost. 1= Provided no information. 2= Provided limited information only on 1 aspect of PASC (either P or A or S or C) 3= Provided detailed information only on 2 aspects of PASC. 4= Provided detailed information only on 3 aspects of PASC. 5= Provided detailed information on all 4 aspects of PASC.		appropriately. There is no hidden cost. 1= Provided no information/Cannot be assessed. 2= Provided information only on 1 aspect of PASC (either P or A or S or C) 3= Provided information only on 2 aspects of PASC. 4= Provided information only on 3 aspects of PASC. 5= Provided information on all 4 aspects of PASC.		
B. *DEVELOPER CREDIBILITY*						
CREDIBILITY	2	How would you rank the credibility and reliability history of the app developer? Please refer to the app description, the app, developer website or T & C statement. Please consider the expertise and reputation of	2	How would you rank the credibility and reliability history of the app developer? Please refer to the app description, the app, developer website or T & C statement. Please consider the expertise and reputation of	**IRR:** Slight Agreement For the top-rated app (ICC, Alpha >0.7), the level of reliability was lower in the domain of *Developer Credibility*. For the items' reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24, and 26 are the question	Scoring options were updated. Points 2 and 3 were moved to screening. Points 4 and 5 were combined and reworded.

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		the developer and their organisation on the following points: 1= Cannot be assessed 2= The name or organisation and contact details of the app developer is NOT available within the app or its supporting information. (e.g. e-mail address, phone number, address, or developers website) 3= The name or organisation and contact details of the app developer is available. However, there is no evidence provided of track record in the specific healthcare domain. 4= The app is either funded, supported, endorsed, or developed by a reputable NGO/institution (hospital/centre, etc.) / university /specialised		the developer and their organisation on the following points: 1= Cannot be assessed 3= Independent organisation but with no evidence provided of track record in the specific healthcare domain. 5= The app is either funded, supported, endorsed, or developed by a reputable NGO/institution (hospital/centre, etc.) / university /specialised commercial business, health funder (verifiable), state/ territory/federal government, or other nationally recognised government bodies or research organisation (e.g. any app backed by Australian Research Council, NHMRC)	with poorer reliability within the domains. The issue of inconsistency can be resolved by offering detailed training for questions Q2-3, 13-15, and 17. **FI:** *Developer Credibility* was noted as an important domain, especially in the unregulated environment. Evaluators should consider the app maker's credibility and the factors driving the app content (i.e., source of funds or who commissioned the app. However, there were opposing views saying that the credibility of the maker of the app should not be the basis of judging a quality app **SDA:** However, the rankings of those who had not previously used health apps differed significantly, ranking *Developer Credibility* as the least important domain.	

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		commercial business, health funder (verifiable) 5= The app is either funded, supported, endorsed, or developed by the state / territory / federal government, or other nationally recognised government bodies or research organisation (e.g. any app backed by Australian Research Council, NHMRC)				
TRANSPARENCY	3	How clearly does the app provide details about the sponsor/developer/affiliations? Please refer to the developer link and/or the app description prior to app downloading. • The app identifies any partner organisations which have been involved in developing the app. • App description provides other app(s) made by the	3	Does the app provide details about the sponsor/developer/affiliations? Please refer to the developer link and/or the app description prior to app downloading. • The app identifies any partner organisations which have been involved in developing the app. • The app developer provide information about	**FB:** 1. If the developer is an independent organisation/company, affiliation should not be applicable and need to count the score for affiliation. 2. Transparency. It was hard to assess the transparency of conflicts of interests as government-backed. 3. What if the developer is a private with experience in clinical practice? 4. Some developers have developed only one app and have provided it under the developer's apps link. so one score should add up if the developer has developed only that app and no other apps.	Point 2 was reworded.

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		developer/partner organization. • Information about the affiliations of the developers is available. • Information is provided on potential commercial interests such as investors and/or any other potential conflicts of interest. 1= Met none of the criteria above. 2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met all of the criteria above.		whether they have made or develop other health apps. • Information about the affiliations of the developers is available. • Information is provided on potential commercial interests such as investors and/or any other potential conflicts of interest. 1= Met none of the criteria above/Cannot be assessed. 2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met all of the criteria above.	5.Q3 has a subjective judgment **IRR**: Slight Agreement For the top-rated app (ICC, Alpha >0.7), the level of reliability was lower in domain of *Developer Credibility* For the items' reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24 and 26.are the question with poorer reliability while within the domains. The issue of inconsistency can be resolved by offering detailed training for questions Q2-3, 13-15 and 17.	
DEVELOPMENT PROCESS	4	To what extent has the app been expertly developed according to the following criteria? Please refer to the developer link noted in the app description prior to app downloading.	4	To what extent has the app been expertly developed according to the following criteria? Please refer to the developer link noted in the app description prior to app downloading.	**FB:** There is an option that needs to say, "could not be assessed." Check typo "demonstrated." End-users were involved in the testing (e.g., interface design involved user participation or been informed by studies of user needs). - Do apps report	Added "could not be assessed as an option."

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		• Relevant experts were involved in the app development and codesign (e.g. clinical, academic) • App has undergone user acceptability and usability testing with details. • End-users were involved in the testing (e.g. interface design involved user participation or been informed by studies of user needs). • Evidence is provided that the app effectiveness has been demonstrated in observational or clinical studies. 1= Met none of the statements above. 2= Met one of the statements above 3= Met two of the statements above 4= Met three of the statements above		• Relevant experts were involved in the app development and codesign (e.g. clinical, academic) • App has undergone user acceptability and usability testing with details. • End-users were involved in the testing (e.g. interface design involved user participation or been informed by studies of user needs). • Evidence is provided that the app effectiveness has been demostrated in observational or clinical studies. 1= Met none of the statements above/ could not be assessed. 2= Met one of the statements above 3= Met two of the statements above 4= Met three of the statements above	on this? I haven't seen it yet. Not sure how to find out **IRR:** Fair agreement For the top-rated app (ICC, Alpha >0.7), the level of reliability was lower in domain *Developer Credibility* **FI:** There needs to be a clear definition of 'Developer' in the framework as confusion can occur depending on various disciplines	

Domain **Sub-domain**	**#**	**Framework Version 2**	**#**	**Framework Version 3**	**Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/** **Focus Interview (FI)/ Survey data analysis (SDA)**	**Actions made to address the comments and results of the analyses**
		5= Met all of the statements above		5= Met all of the statements above		
C. ***CONTENT VALIDITY***						
REPORTED REFERENCE(S)	5	Is the health information within the app and or app description appropriately referenced? • App content is partially referenced from credible health sources (e.g. official guidelines, peer-reviewed systematic reviews etc.) • App content is fully referenced from credible health sources (e.g. official guidelines, peer-reviewed systematic reviews etc.) • Health information displays dates of when the health information was created or modified in the app • Medical experts/clinical guidelines, authors, contributors, editors, or committee name are reported	5	Is the health information within the app and or app description appropriately referenced? • App content is partialy referenced from credible health sources (e.g. official guidelines, peer-reviewed systematic reviews etc.) • App content is fully referenced from credible health sources (e.g. official guidelines, peer-reviewed systematic reviews etc.). (If chosen please add one point, assuming that full reference is better than partial reference) • App developer websites/app description provide dates of when the health information was created or modified in the app • Medical experts/clinical guidelines, authors,	**FB:** point 1 and 2 as 2 is better then one Health information displays dates of when the health information was created or modified in the app - it might just be the apps I'm trying to assess, but I don't really see dates for things like chatbots. There are recent publications which I didn't count as you're not sure if the findings feedback into modifications. From website info it's implied they're regularly updating, but no dates are provided. Partial of full reference, should it be adding one additional score when the app has full ref. Information that can only be accessed with in-app purchases has not been reviewed for reference and content info. Information content can't be assessed due to in-app purchases	Changed the wordings to increase agreement. Add option “Cannot be assessed.”

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		1= Met none of the statements above. 2= Met one of the statements above 3= Met two of the statements above 4= Met three of the statements above 5= Met all of the statements above		contributors, editors, or committee name are reported 1= Met none of the statements above/Cannot be assessed. 2= Met one of the statements above 3= Met two of the statements above 4= Met three of the statements above 5= Met all of the statements above	**IRR:** Fair agreement **FI:** Participants made recommendations to check whether app content was based on an evidence-based source. The most reliable evidence is randomised-controlled trials followed by clinical guidelines, government health guidelines, expert opinions, or standard healthcare practices. *Content/Information Validity* appears to be the most important domain. **SDA:** Based on weighted average rankings, 'Content validity' was seen as the most important domain by all the participants in general and by those who had previously used health apps	
PRESENTATION OF APP CONTENT	6.	How is the information/content presented within the app? • The information is comprehensive and concise	6	How is the information/content presented within the app? • The information is comprehensive and concise	**FB:** the app had videos as content, but it was not audible and, therefore, could not access the content. The second and third points again should be just "information," removing "health."	"Health was removed."

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		(i.e. containing basic facts about the nature of the condition, treatment or medications or procedure). • The health information is clearly presented (i.e avoidance of foreign or overly technical or clinical terms). • The health information is well written (i.e free of grammatical errors, objective without emotional overtones). • The presentation of the information is engaging and in logical manner 1= Met none of the statements above/cannot be assessed. 2= Met one of the statements above 3= Met two of the statements above 4= Met three of the statements above 5= Met all of the statements above		(i.e. containing basic facts about the nature of the condition, treatment or medications or procedure). • The information is clearly presented (i.e avoidance of foreign or overly technical or clinical terms). • The information is well written (i.e free of grammatical errors, objective without emotional overtones). • The presentation of the information is engaging and in logical manner 1= Met none of the statements above/cannot be assessed. 2= Met one of the statements above 3= Met two of the statements above 4= Met three of the statements above 5= Met all of the statements above	**IRR:** Slight Agreement Opportunity for improvement is detected if removing questions 6,16,17,18, and 20. However, it needs to be confirmed by analysis per app and per domain approaches. **FI:** Personalisation of content should also be reviewed as this could provide a continuity of care for the patient healthcare journey. **SDA:** Based on weighted average rankings, *Content/Information Validity* was seen as the most important domain by all the participants in general and by those who had previously used health apps.	

Domain **Sub-domain**	**#**	**Framework Version 2**	**#**	**Framework Version 3**	**Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/** **Focus Interview (FI)/ Survey data analysis (SDA)**	**Actions made to address the comments and results of the analyses**
QUALITY AND EFFECTIVNESS OF INFORMATION	7	How is the evaluation of safety and quality of the app and information presented? Please refer to the app description, notifications within the app or T&C. A. Safety and quality information on the app is provided, AND appears appropriate and relevant to the target audience. B. A statement is provided about how often the health content will be reviewed and updated according to the most current clinical evidence. C. The app content is supported by a lower grade of evidence (e.g. clinical or academic expert advice, government documents, clinical guidelines, text books). D. The app content is supported by a higher grade of evidence (e.g.	7	What supporting evidence is presented for the quality and effectiveness of the information in the app or referenced by the app? Please refer to the app description, notifications within the app or T&C. Please check some of the references directly to validate. A. Quality and effectiveness of information on the app is provided, AND appears appropriate and relevant to the target audience. B. A statement is provided about how often the health content will be reviewed and updated according to the most current clinical evidence. C. The app content is supported by a lower grade of evidence (e.g. clinical or academic expert advice, government documents,	**FB:** (needs a greater point for dot point 4). the authors/institutions seem legitimate and state the app is backed by research. There is a page with "research" listed. But no links or information with results about research is listed. Shall evaluators check those with claims using different search engines? some clarity around the wording of "the evaluation of the safety and quality of the app and information presented" (sorry can't remember the question number). Is this specifically referring to the reliability/trustworthiness of the content of the app? **IRR:** Slight Agreement **SDA:** Based on weighted average rankings, *Content/Information Validity* was seen as the most important domain by all the participants in general and by those who had previously used health apps.	Sub-domain was updated from “safety and quality of information” to “quality and effectiveness of information” Changed the wordings of the questions and points.

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		scientific evidence from descriptive, observational or experimental study that has been published in a journal, systematic review, health technology assessment). Note: If selecting D, please also count statement C. 1= Met none of the statements above. 2= Met one of the statements above 3= Met two of the statements above 4= Met three of the statements above 5= Met all of the statements above		clinical guidelines, text books). D.The app content is supported by a higher grade of evidence (e.g. scientific evidence from descriptive, observational or experimental study that has been published in a journal, systematic review, health technology assessment). Note: If selecting D, please also count statement C. 1= Met none of the statements above/Cannot be assessed. 2= Met one of the statements above 3= Met two of the statements above 4= Met three of the statements above 5= Met all of the statements above		

Domain **Sub-domain**	**#**	**Framework Version 2**	**#**	**Framework Version 3**	**Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/** **Focus Interview (FI)/ Survey data analysis (SDA)**	**Actions made to address the comments and results of the analyses**
CONTEMPORARY INFORMATION USED WITHIN THE APP	8	When was the safety and quality information/content within the app last updated? 1 = more than 18 months/no information 2 = 12 to 18 months 3 = 6 to less than 12 months 4 = 3 to less than 6 months 5 = under 3 months	8	How recent is the information presented within the app? (note: What is the publication date of the most recently cited evidence or reference)? 1 = no information provided or no references 2 = greater than 5 years 3 = less than 5 years 4 = less than 3 years 5 = less than 1 year	**FB:** most apps provide the history of updates, but information update is very rare and mostly bug fixes. need to say where to find the information update details (Version history and look for any information updates) **IRR:** Slight Agreement **SDA:** Based on weighted average rankings, Content/*Information Validity* was seen as the most important domain by all the participants in general and by those who had previously used health apps.	Sub-domain was updated from "safety and quality -frequency of update/checking the accuracy of information" to "contemporary information used within the app.**"** Wordings were updated and revised.
D. ***PRIVACY/SECURITY/ ETHICS/LEGAL***						
DATA COLLECTION AND USE	9	How well was the data collection process and data usage described? Please refer to T&C, Privacy Policy statements, the app description or within the app. • The process, use, handling of personal data was described.	9	How well was the data collection process and data usage described? Please refer to T&C, Privacy Policy statements, the app description or within the app. • The process, use, handling of personal data was described.	**FB:** Need a question that asks if the app could be used without providing personal details or registration Does this need to be a function within the app? There isn't a button to download data, but the privacy policy says you can request it **IRR:** Slight Agreement	*Privacy/Security/Ethics/ Legal* section was not updated as it appeared that this is the hardest and weakest area of the framework. At the same time, this was noted as one of the most

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		• The purpose of the data collection was described. • Health app describes which personal information items are collected and recorded. • App describes options available to users to manage their personal data and privacy. 1= Met none of the statements above. 2= Met one of the statements above 3= Met two of the statements above 4= Met three of the statements above 5= Met all of the statements above		• The purpose of the data collection was described. • Health app describes which personal information items are collected and recorded. • App describes options available to users to manage their personal data and privacy. 1= Met none of the statements above/Cannot be assessed. 2= Met one of the statements above 3= Met two of the statements above 4= Met three of the statements above 5= Met all of the statements above	The analysis indicated that the domain *Privacy/Security/Ethics/Legal* has the poorest average ICC result in the poorly rated app (n=10/14) In consideration for items within each domain, all items within *Privacy/Security/Ethics/Legal* appeared to have unreliable items toward showing a poor agreement for this domain (removing items will lead to an increment in the alpha score: Alpha toward 0.7) For the items reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24, and 26 are the question with poorer reliability within the domains. A further refinement might be required for questions 9-12 and 24-27, which will be confirmed in approach 3 (analysis for the consistency within domains). A further refinement might be required for questions 9-12 and 24-27.	important domains in the focus interview. Further investigation is needed.

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
					FI: Participants also suggested checking whether the apps permit end-users to delete their data.	
PRIVACY POLICY AND TERMS AND CONDITIONS	10	ow do the T&C address privacy? Please refer to T&C/privacy statements or within the app. • T&C/privacy statement is available within the app itself or in the app store. • It is readable to the general adults (grade level below 8 using F-K reading scale). (You may need to copy and paste the T&C to a Word document and run the readability statistics: https://support.microsoft.com/en-us/topic/get-your-document-s-readability-and-level-statistics-85b4969e-e80a-4777-8dd3-f7fc3c8b3fd2)	10	How do the T&C address privacy? Please refer to T&C/privacy statements or within the app. • T&C/privacy statement is available within the app itself or in the app store. • It is readable to the general adults (grade level below 8 using F-K reading scale). (You may need to copy and paste the T&C to a Word document and run the readability statistics: https://support.microsoft.com/en-us/topic/get-your-document-s-readability-and-level-statistics-85b4969e-e80a-4777-8dd3-f7fc3c8b3fd2)	**FB:** 1. The privacy statement web page did not allow copying in order to check readability. 2. It is very hard to do on a small iPhone 3. It explicitly states that they are following privacy standards in Australia (Australian Privacy Act) and/or European General Data Protection Regulation (GDPR). - Privacy policy states they follow GDPR but only for people within Europe. So, is it a yes, or no as not applicable to Australian users? 4. Readability (change scale): https://link.springer.com/chapter/10.1007/978-3-030-49669-2_22 5. Training needs to be done for reading T&C and privacy 6. Last point. if any of these items are present. 7. policy statement of all the apps evaluated so far has readability higher than grade 8, which may suggest that	Same as above

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		Note: After setting up the readability statistics in MS Word (refer to the provided link), copy and paste the T&C/privacy policy to the word document. Press "F7" then click on the “total suggestion” and ignore all those suggestion(s). The readability statistics should pop after these steps and refer to the Flesch-Kincaid Grade level to assess this criteria). • It explicitly states that they are following privacy standards in Australia (Australian Privacy Act) and/or European General Data Protection Regulation (GDPR). • It has statements related to security protection including confidentiality, safe environment, conditions of services in the cloud/storage location,		Note: After setting up the readability statistics in MS Word (refer to the provided link), copy and paste the T&C/privacy policy to the word document. Press "F7" then click on the “total suggestion” and ignore all those suggestion(s). The readability statistics should pop after these steps and refer to the Flesch-Kincaid Grade level to assess this criteria). • It explicitly states that they are following privacy standards in Australia (Australian Privacy Act) and/or European General Data Protection Regulation (GDPR). • It has statements related to security protection including confidentiality, safe environment, conditions of services in the cloud/storage location,	using the F-K grade scale might not be the best option for checking readability. 8. Grade level of readability should be adjusted to https://link.springer.com/chapter/10.1007/978-3-030-49669-2_22. 9. How about those Privacy/T&C not written in English? 10. Privacy should include information related to children. 11. A guidebook for reading through t&cs and privacy policies would be good - what you could ctrl+f for. I found reading this the most tedious and time-consuming part of testing the app and completing the questions. 12. when talking about privacy policies, it's unclear whether apps have to meet all of the criteria (confidentiality, safe environment, conditions of services in the cloud/storage location, personal information and safety standards, offshore transmission of unidentifiable data, and retention of unencrypted data) or any number of these to be marked as "met."	

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		personal information and safety standards, offshore transmission of identifiable data, and retention of unencrypted data. 1= Met none of the statements above. 2= Met one of the statements above 3= Met two statements above 4= Met three statements above 5= Met all the statements above		personal information and safety standards, offshore transmission of indentifiable data, and retention of unencrypted data. 1= Met none of the statements above/Cannot be assessed. 2= Met one of the statements above 3= Met two statements above 4= Met three statements above 5= Met all the statements above	**IRR**: Slight Agreement The analysis indicated that *Privacy/Security/Ethics/Legal* has the poorest average ICC result in the poorly rated apps (n=10/14) In consideration for items within each domain, all items within Domain D appeared to have unreliable items toward showing a poor agreement for this domain (removing items will lead to an increment in the alpha score: Alpha toward 0.7) For the items reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24, and 26 are the question with poorer reliability within the domains. A further refinement might be required for questions 9-12 and 24-27, which will be confirmed in approach 3 (analysis for the consistency within domains). A further refinement might be required for questions 9-12 and 24-27.	

Domain **Sub-domain**	**#**	**Framework Version 2**	**#**	**Framework Version 3**	**Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/** **Focus Interview (FI)/ Survey data analysis (SDA)**	**Actions made to address the comments and results of the analyses**
					FI: Privacy/security/ethical/legal appeared to be the most important domain. Recommendation: Priority should be given to *Privacy/Security/Ethics/Legal* domain and *Content/Information Validity* domains. If we consider the pyramid or weighted average approach, these two domains should be given priority. The clarity of purpose was also ranked as the third most important by the participants.	
DECLARATION OF RISK(S)	1 1	Does the app clearly display a declaration of risk? Please refer to the app description, notifications within the app or T&C/ user license (or contract) agreement. • There is a statement that the app should not replace the advice of a health professional OR the app advises you to consult a doctor before starting or using the app.	1 1	Does the app clearly display a declaration of risk? Please refer to the app description, notifications within the app or T&C/ user license (or contract) agreement. • There is a statement that the app should not replace the advice of a health professional OR the app advises you to consult a doctor before starting or using the app.	**FB:** (acknowledgment of T&C are not about 18 years old) It's useful to have questions like this right at the start because I explore the app first and answer a lot of other questions and then can't remember exactly and have to uninstall or log out to double-check. Most do anyway. [The user needs to read and accept the "general terms and conditions of use" before starting to use the app] **IRR: Fair agreement**	Same as above

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		• The app provides information related to the potential risks, errors or side-effects resulting from its use. • The app provides a third-party safety contact link or phone number available in Australia (e.g. lifeline) in the event of distress. • The user needs to read and accept the "general terms and conditions of use" before starting to use the app". 1= Met none of the statements above. 2= Met one of the statements above 3= Met two statements above 4= Met three statements above 5= Met all the statements above		• The app provides information related to the potential risks, errors or side-effects resulting from its use. • The app provides a third-party safety contact link or phone number availalble in Australia (e.g. lifeline) in the event of distress. • The user needs to read and accept the "general terms and conditions of use" before starting to use the app". 1= Met none of the statements above/Cannot be assessed. 2= Met one of the statements above 3= Met two statements above 4= Met three statements above 5= Met all the statements above	The analysis indicated that *Privacy/Security/Ethics/Legal* has the poorest average ICC result in the poorly rated app (n=10/14) In consideration for items within each domain, all items within *Privacy/Security/Ethics/Legal* appeared to have unreliable items toward showing a poor agreement for this domain (removing items will lead to an increment in the alpha score: Alpha toward 0.7) For the items' reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24, and 26 are the question with poorer reliability within the domains. A further refinement might be required for questions 9-12 and 24-27, which will be confirmed in approach 3 (analysis for the consistency within domains). A further refinement might be required for questions 9-12 and 24-27.	

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
DATA STORAGE AND SECURITY	1 2	How is data storage, handling and security described? • A statement is provided about where the data are stored (e.g. locally on device or in a vendor cloud store). • Users have the ability to control sharing of data with vendor or other parties. • App describes data security features such as single- or multiple-user authentication and authorization/ data signature and tampering, password protection, age verification process, encrypted in transit between the device and any developer host storage. • App can be used without the collection of personally identifiable information. 1= Met none of the criteria above.	1 2	How is data storage, handling and security described? • A statement is provided about where the data are stored (e.g. locally on device or in a vendor cloud store). • Users have the ability to control sharing of data with vendor or other parties. • App describes data security features such as single- or multiple-user authentication and authorization/ data signature and tampering, password protection, age verification process, encrypted in transit between the device and any developer host storage. • App can be fully used without the collection of personally identifiable information/registration.	**FB:** I still think this is problematic. If the data is only stored on the device or can be used without collecting info this needs to be scored at a top-level - not just as one of several criteria if the app does not store or collect personal data, then this would score zero, which is unfair **IRR:** Fair agreement The analysis indicated that *Privacy/Security/Ethics/Legal* has the poorest average ICC result in the poorly rated app (n=10/14) In consideration for items within each domain, all items within *Privacy/Security/Ethics/Legal* appeared to have unreliable items toward showing a poor agreement for this domain (removing items will lead to an increment in the alpha score: Alpha toward 0.7) For the items' reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24, and 26 are the question	Same as above

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met all of the criteria above.		1= Met none of the criteria above/Cannot be assessed. 2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met all of the criteria above.	with poorer reliability within the domains. A further refinement might be required for questions 9-12 and 24-27, which will be confirmed in approach 3 (analysis for the consistency within domains). A further refinement might be required for questions 9-12 and 24-27. FI: Evaluators should check if the data are stored in Australia as there are issues related to ownership and control of data when stored offshore	
E. *TECHNICAL FEATURES AND SUPPORT*						
SYSTEM DEFECTS/ ERROR/ BUGS	13	Is the app free of obvious defects/errors or bugs? You will need to test the following by opening and using the app. • The functionalities of the app (such as saving data, performing calculations, playing music, data entry) are operative and without time lagging or crashing of the app.	13	Is the app free of obvious defects/errors or bugs? You will need to test the following by opening and using the app. • The functionalities of the app (such as saving data, performing calculations, playing music, data entry) are operative and without time lagging or crashing of the app.	**IRR**: Slight Agreement For the items' reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24, and 26 are the question with poorer reliability within the domains. The issue of inconsistency can be resolved by offering detailed training for questions Q2-3, 13-15, and 17. **FI:** *Technical features and Support* are one of the least important	No changes

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		• The components of the app (buttons/menus/ textboxes) are labelled correctly, operative, accessible and without time delay (apart from Internet lag which is outside control of the app). • Navigation between different parts of the app and data entry (if any) is smooth - there are no times that the user is frozen on a particular screen or that the app crashes. • Moving between screens is logical/accurate/appropriate/ with clear navigation logic and ability to go directly to an app Home or Help page (or similar) in one step. 1= Met none of the criteria above. 2= Met one of the criteria above.		• The components of the app (buttons/menus/ textboxes) are labelled correctly, operative, accessible and without time delay (apart from Internet lag which is outside control of the app). • Navigation between different parts of the app and data entry (if any) is smooth - there are no times that the user is frozen on a particular screen or that the app crashes. • Moving between screens is logical/accurate/appropriate/ with clear navigation logic and ability to go directly to an app Home or Help page (or similar) in one step. 1= Met none of the criteria above/Cannot be assessed. 2= Met one of the criteria above.	**SDA:** However, the rankings of those who had not previously used health apps differed significantly, ranking *Technical features and Support* as the most important.	

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met all of the criteria above.		3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met all of the criteria above.		
APP MAINTENANCE	14	How frequently is the app being maintained to ensure that the app is free from defects, errors, or bugs? When was the last update: 1= Not stated or the application was revised over 18 months ago. 2= The application was revised 13-18 months ago. 3= The application was revised 7-12 months ago. 4= The application was revised 3-6 months ago. 5= The application has been revised less than 3 months ago.	14	How frequently is the app being maintained to ensure that the app is free from defects, errors, or bugs? When was the last update: 1= Not stated or the application was revised over 18 months ago/Cannot be assessed. 2= The application was revised 13-18 months ago. 3= The application was revised 7-12 months ago. 4= The application was revised 3-6 months ago. 5= The application has been revised less than 3 months ago.	**IRR:** Substantial agreement Disagreement was only shown in one question (Q14), which is reasonable. Q14 was seen as having a high agreement in top-rated apps and poor agreement in the poorly rated app, indicating poor agreement in the poorest rated app can be solved by detailed training. The issue of inconsistency can be resolved by offering detailed training for questions Q2-3, 13-15, and 17. **FI:** *Technical features and Support* are one of the least important **SDA:** However, the rankings of those who had not previously used health apps differed significantly, ranking *Technical features and Support* as the most important.	No changes

Domain **Sub-domain**	**#**	**Framework Version 2**	**#**	**Framework Version 3**	**Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/** **Focus Interview (FI)/ Survey data analysis (SDA)**	**Actions made to address the comments and results of the analyses**
CUSTOMER SERVICES	15	Is there any mechanism whereby users can report issue(s) and provide feedback related to the app? (Note: some apps may ask for feedback when you quit the app) • Feedback can be provided by the Appstore or playstore. • There is a statement within the app about the developer's commitment to address problems reported to them (e.g. timescales to respond, commitment to eradicate reported bugs and faults). • There is a mechanism to contact technical support (e.g. e-mail or phone number is given to contact the technical support team). • The developer provided an in-app means (e.g. chat, form) for users to provide feedback on the quality and usability of the app.		Deleted	**IRR:** Slight Agreement For the items' reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24, and 26 are the question with poorer reliability within the domains. The issue of inconsistency can be resolved by offering detailed training for question **FI:** *Technical features and Support* are one of the least important **SDA:** However, the rankings of those who had not previously used health apps differed significantly, ranking *Technical features and Support* as the most important.	Deleted to improve agreement

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		1= Met none of the criteria above. 2= Met one of the criteria above. 3=Met two of the criteria above. 4= Met three of the criteria above. 5= Met all the criteria above.				
F. *ACCESSIBILITY*						
LANGUAGE AVAILABILITY	16	Does the app include relevant languages for its targeted group and how many languages spoken in Australia does the app offer? (Top languages spoken at home in Australia are the following Mandarin, Arabic, Cantonese, Vietnamese, Italian, Greek, Hindi, Spanish, Punjabi according to ABS 2016 1= App is only available in English language	15	Does the app include relevant languages for its targeted group and how many languages spoken in Australia does the app offer? (Top langugages spoken at home in Australia are the following Mandarin, Arabic, Cantonese, Vietnamese, Italian, Greek, Hindi, Spanish, Punjabi according to ABS 2016 1= App is only available in English language	**FB:** point three should read six and above **IRR:** Moderate agreement Opportunity for improvement is detected if removing questions 6,16,17,18, and 20. However, it needs to be confirmed by analysis per app and per domain approaches. The analysis indicated that *Accessibility* has a poor average ICC result in the poorly rated app (n= 9/14).	No changes as moderate agreement

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		2= App supports the targeted group and two ABS top 10 languages 3= App supports the targeted group and three to five ABS top 10 languages 4= App supports the targeted group and five six to seven ABS top 10 languages 5= App supports the targeted group and eight or more ABS top 10 languages OR the app supports any Australian Aboriginal language		2= App supports the targeted group and two ABS top 10 languages 3= App supports the targeted group and three to five ABS top 10 languages 4= App supports the targeted group and six to seven ABS top 10 languages 5= App supports the targeted group and eight or more ABS top 10 languages OR the app supports any Australian Aboriginal language	For top-rated app (ICC, Alpha >0.7), the level of reliability was lower in *Accessibility*	
AVAILABILITY ACROSS PLATFORMS	17	How compatible is the app across platforms? Please refer to the developer link noted in the app description • The app is usable in iOS (Apple). • The app is usable in Android (Google, Samsung).	16	Which of the following platforms does the developer states the app is available? Please refer to the app description or developer link. • The app is available only in one of iOS (Apple) or Android (Google, Samsung)	**FB:** '-I am only testing on one device. As far as I can tell, there is no way for me to assess which other devices the apps are compatible across. -If we are using one device (in my case, an iPhone since shifting from Android) – how can the reviewer answer the question? I think this is a question that we, as the screening team, can best determine and not require the evaluator to assess.	Sub-domain was updated from "peripheral devices" to "availability across platforms." The question was reworded to improve the agreement.

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		• The app is usable in other smartphone platform (Blackberry OS (Research in Motion), Windows OS (Microsoft), Symbian OS (Nokia) or others). • The app supports and optimize smartphone, tablets/ "slate" devices. • The app is compatible with different Versions of operating systems (e.g. iOS 9.0 or higher OR Android 1.0 or higher). 1= Met one of the criteria above. 2= Met two of the criteria above. 3= Met three of the criteria above. 4= Met four of the criteria above. 5= Met all of the criteria above.		• The app is available in both iOS (Apple) and Android (Google, Samsung) • The app is available in other smartphone platform (Blackberry OS (Research in Motion), Windows OS (Microsoft), Symbian OS (Nokia) or others). • The app supports and optimize smartphone, tablets/ "slate" devices. • The app is compatible with different Versions of operating systems (e.g. iOS 9.0 or higher OR Android 1.0 or higher). 1= Met one of the criteria above/Cannot be assessed. 2= Met two of the criteria above. 3= Met three of the criteria above. 4= Met four of the criteria above. 5= Met all of the criteria above.	The app is compatible with different versions of operating systems (e.g., iOS 9.0 or higher OR Android 1.0 or higher). - I was not sure whether the examples were minimum requirements for how far back Versions needed to be compatible. **IRR:** Poor Agreement An opportunity for improvement is detected if removing questions 6,16,17,18 and 20. However, it needs to be confirmed by analysis per app and per domain approaches. The analysis indicated that *Accessibility* has a poor average ICC result in the poorly rated app (n= 9/14). For top-rated app (ICC, Alpha >0.7), the level of reliability was lower in Domain *Accessibility* For the items' reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24 and 26 are the question with poorer reliability. The	

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
					inconsistency issue can be resolved by offering detailed training for questions Q2-3, 13-15, and 17.	
OFFLINE FUNCTIONALITY	18	Does the app need an active connection to the Internet for necessary app functions to work? 1= The app works only with an Internet connection. 3= Without Internet connection, the app has some restrictions (e.g. it displays static data such as contacts and/or accepts user input and/or provides user interactions) 5= The app is fully functional without Internet.	17	Does the app need an active connection to the Internet for necessary app functions to work? 1= The app works only with an Internet connection. 3= Without Internet connection, the app has some restrictions (e.g. it displays static data such as contacts and/or accepts user input and/or provides user interactions, requires Internet for initial registration or downloads, but is usable offline afterwards) 5= The app is fully functional without Internet.	**FB:** This needs a bit of tuning. For example, "Anxiety Reliever" requires the Internet to download content – after which it can be used offline. Perhaps a fourth option says: "Requires Internet for initial registration or downloads, but is usable offline afterwards" How about apps that really require an internet connection (e.g., peer support app)? Suggest different overall scoring for this item Question related to online/offline is problematic **IRR:** Fair agreement Opportunity for improvement is detected if removing questions 6,16,17,18, and 20. However, it needs to be confirmed by analysis per app and per domain approaches.	Added examples. Further investigation to differentiate apps that do and do not allow real-time communication (e.g., peer support apps)

Domain **Sub-domain**	**#**	**Framework Version 2**	**#**	**Framework Version 3**	**Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/** **Focus Interview (FI)/ Survey data analysis (SDA)**	**Actions made to address the comments and results of the analyses**
					The analysis indicated that *Accessibility* has a poor average ICC result in the poorly rated app (n= 9/14). For top-rated app (ICC, Alpha >0.7), the level of reliability was lower in Domain *Accessibility*	
VULNERABLE POPULATION	19	What support does the app offer for vulnerable populations? e.g. to improve *Accessibility* and acceptability to target audiences who may have limited digital and or health literacy, cultural barriers, and visual or hearing impairment (DOES NOT include languages, see Q17) • The app provides support options for users with visual impairment (e.g. changing font sizes or colour, voice overs). • The app provides support options for users with	18	What support does the app offer for vulnerable populations? e.g. to improve *Accessibility* and acceptability to target audiences who may have limited digital and or health literacy, cultural barriers, and visual or hearing impairment (DOES NOT include languages, see Q17) • The app provides support options for users with visual impairment • The app provides support options for users with hearing difficulties	**FB**: The iOS (and Android) systems provide a LOT of *Accessibility* features now. So, we could specify in the text that we refer to *Accessibility* features OVER and ABOVE that are provided by the device. For example, the BBC News app (which I use every day) drives me crazy because of its tiny text. It has an in-app text size thing that just points you to the phone text resize feature. But the app does NOT respect the phone auto text resize feature (unlike the ABC news app, which does) – and so the BBC News App gets a big fail for *Accessibility* from me even though they claim to support it by providing the menu option for text size. **IRR:** Slight Agreement	We tried to include the q's or criteria for vulnerable pop, but we found that we only have a slight agreement and have a complicated overlap between the "apps themselves" and the IOS features. An area that needs further development

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		hearing difficulties (e.g. with sound amplifiers or providing text in place of voice/sound). • The app provides support options to target audiences who may have limited digital and/or health literacy (i.e apps programmed with computer-animated characters, text, and graphics) • The app provides support options for users with other specific *Accessibility* barriers (e.g. comorbidities, religious dietary requirements etc.) 1= Met none of the criteria above. 2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above.		1= Met none of the criteria above/Cannot be assessed. 3= Met one of the criteria above. 5= Met two of the criteria above.	The analysis indicated that *Accessibility* has a poor average ICC result in the poorly rated app (n= 9/14). For the top-rated app (ICC, Alpha >0.7), the level of reliability was lower in domain *Accessibility* **FI:** Look beyond the language or online and offline functionality and review how the health apps extend or likely bridge the care for people with a disability or work with vulnerable population groups.	

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		5= Met all of the criteria above.				
COST	20	What does the app's basic plan cost you per year? (excluding in-app purchases) 1= Estimated annual cost > $100 2= Estimated annual cost between $11-$99 3= Estimated annual cost < $10 4= The app is "free" but is supported by adverts or has premium higher costs features that need to be purchased or subscribed 5= The app is "free" without advertisements, no subscription required	19	What does the app's basic plan cost you per year? (excluding in-app purchases) Note: where the in-app pricing appears in USD please convert it to AUD. 1= Estimated annual cost > AUD 100 2= Estimated annual cost between AUD 11-99 3= Estimated annual cost < AUD 10 4= The app is "free" but is supported by adverts or has premium higher costs features that need to be purchased or subscribed 5= The app is "free" without advertisements, no subscription required, no premium features.	**FB: '-** should add something (the app is free but endorsement to buy a specific product ex: training program) -many of the apps appear to be US and the costs are $US. How do you want to work this? I think we should convert to AUD. So maybe we could just provide a bit of text saying, "where the in-app pricing appears in USD, for guidance, presently, $1AUD is worth $0.75USD it is free - but they harvest your social media data and require you to sign-up **IRR:** Moderate agreement Opportunity for improvement is detected if removing questions 6,16,17,18, and 20. However, it needs to be confirmed by analysis per app and per domain approaches.	Added notes. This question is only looking at the financial cost. Unnecessary data collection is already penalised in other domains.

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
					The analysis indicated that *Accessibility* has a poor average ICC result in the poorly rated app (n= 9/14). For top-rated app (ICC, Alpha >0.7), the level of reliability was lower in *Accessibility* **SDA**: Majority are not willing to pay	
G. *INTEROPERABILITY*						
INTEROPERABILITY	21	How does the app share and transfer the data? • The app allows data sharing to healthcare professionals. • The app allows data integration with e-health apps (e.g. Apple Health, Google Fit, Samsung Health). • The app allows data integration with electronic health records. (e.g. My Health Record) • The app has the ability to import/export data under user management.	20	How does the app share and transfer the data? • The app allows data sharing to healthcare professionals. • The app allows data integration with e-health apps (e.g. Apple Health, Google Fit, Samsung Health). • The app allows data integration with electronic health records. (e.g. My Health Record) • The app has the ability to import/export data under user management.	**IRR:** Fair agreement **FI:** *Interoperability* is one of the least important **SDA:** Based on weighted average rankings, *Interoperability* was ranked as the least important domain.	No changes

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		1= Cannot be assessed or met none of the criteria 2= Met one of the criteria above. 4= Met two of the criteria above. 5= Met three of the criteria above. 5= Met all of the criteria above.		1= Cannot be assessed or met none of the criteria 2= Met one of the criteria above. 4= Met two of the criteria above. 5= Met three of the criteria above. 5= Met all of the criteria above.		
H. *USER ENGAGEMENT*						
GAMIFICATION or ENTERTAINMENT/ GOAL or TARGET SETTING/ SELF-AWARENESS/ BEHAVIOURAL CHANGE/ SELF-MANAGEMENT/ TIMELY FEEDBACK/ RESULT ANALYSIS	2 2	Does the app use any strategies to improve *User Engagement*? • The app applies gamification techniques (such as digital rewards, prizes, leader boards, badges, or competitions). • The app allows users to set goals to encourage striving and persisting behaviour (e.g. follow-up prompts, reminders) • The app supports users in social connection, sharing and communication (e.g. community discussion,	2 1	Does the app use any strategies to improve *User Engagement*? Please note that we are not assessing the effectiveness of different app features, rather we are only looking if the apps have any of the following features. • The app applies gamification techniques (such as digital rewards, prizes, leader boards, badges, or competitions). • The app allows users to set goals to encourage striving and persisting	**FB:** 'useful' word can be removed from the option as it seems objective. How will we be able to measure the extent of gamification. Some apps are a 30-day "course" which offers different rewards (gamification styles etc.) after different tasks are completed. If we use the apps only for 10 minutes, we may miss these and provide a rating accordingly, which is not accurate. **IRR:** Moderate agreement **SDA**: Respondents reported that their frequency of app use reduced overtime for all three apps. All the	Useful was removed in the Sub-domain name and dot # 4 points. Added notes as well.

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
SOCIAL SUPPORT/ SHARING/ COMMUNICATION		share achievements, social media groups or networks, forums) • The app provides useful analyses and feedback on performance, behaviours, activities, scores, or tests to the user. 1= Met none of the criteria above. 2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met all of the criteria above.		behaviour (e.g. follow-up prompts, reminders) • The app supports users in social connection, sharing and communication (e.g. community discussion, share achievements, social media groups or networks, forums) • The app provides analyses and feedback on performance, behaviours, activities, scores, or tests to the user. 1= Met none of the criteria above/Could not be assessed 2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met all of the criteria above.	apps were rated as easy to use and easy for the user to learn how to use the app, and navigation was convenient moving between screens.	

Domain **Sub-domain**	**#**	**Framework Version 2**	**#**	**Framework Version 3**	**Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/** **Focus Interview (FI)/ Survey data analysis (SDA)**	**Actions made to address the comments and results of the analyses**
USER RATINGS/ DOWNLOADS	2 3	As proxy to *User Experience*: What was the user star rating? Please only refer to the market place available in your smartphone such as Apple app store or Google play store. 1= 0-1 star rating 2= 1.1 to 2.0 star rating 3= 2.1 to 3.0 start rating 4= 3.1 to 4.0 star rating 5= 4+ star rating	2 2	As proxy to *User Experience*: What was the user star rating? Please only refer to the market place available in your smartphone such as Apple app store or Google play store. 1= No rating available/Cannot be assesed 2= 0 to 2.0 star rating 3= 2.1 to 3.0 start rating 4= 3.1 to 4.0 star rating 5= 4+ star rating	**FB**: (should account for the number of raters) There were not enough reviews for it to have an assigned star rating. Only one review was visible, which I used There were no reviews, so it had not been rated. So giving it a Zero was not the same as saying "no reviews". Also, the App website is now deregistered. It would be useful to have that as a screening question so any further evaluation should cease. I also think the question on star rating will have a bit of a ceiling effect, with most apps falling above 3.5* or 4*, and 4+ won't lead to much differentiation amongst high-performing apps. **IRR:** Almost perfect **FB:** *User Engagement* could be reviewed by checking the metrics or analytics of health apps, such as the number of downloads, prevalence of end-user visits, the frequency or	Add another question asking for the number of people who rated the app/the number of downloads. (look at the frequency of the number of raters for the cutoff).

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
					constancy of usage, mobile traffic, and the attrition rate of end-users. **SDA**: Nine out of ten respondents who used 'Smiling Mind' recommended this app to other members. Similarly, eight out of 11 respondents recommended 'Headspace' to other members. The two respondents who used the 'Clear Fear' also recommended this app.	
I. *VALUE*						
PERCEIVED *VALUE*	24	What is the perceived *Value* of the app? • The app is useful for enhancing health or monitoring/mitigating/evaluating a disease, or developing a skill. • The app contributes to the management and/or facilitation of a relevant health care service. (e.g. the app facilitates the service such as enabling book/reschedule/appointm		Deleted	**FB**: I think we should word: "What is the *Value* of the app that you believe an intended end-user would perceive? Rather than: What is the perceived *Value* of the app? should have a point of "cannot be assessed." **IRR:** Slight Agreement For the items' reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24, and 26 are the question	Removed as this is highly subjective

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		ent to the desired service, or the app allows user to communicate with their health providers) • The app provides information about a disease, or health aspect. • The app seeks to change attitudes and or behaviours (awareness, knowledge, attitudes, intention to change, behaviour change, health seeking) 1= Met none of the criteria above. 2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met all of the criteria above.			with poorer reliability within the domains. A further refinement might be required for questions 9-12 and 24-27, which will be confirmed in approach 3 (analysis for the consistency within domains). A further refinement might be required for questions 9-12 and 24-27. **FI:** *Value* domain was said to be highly important, but there is no pragmatic approach discussed on how to evaluate it. Several domains, such as *Clarity of Purpose of the App*, *User Engagement*, *User Experience* and *Accessibility* domains are connected to the *Value* domain. We can recommend that these domains can be reviewed as a proxy of the *Value* domain.	
J. *USER EXPERIENCE*						

Domain **Sub-domain**	**#**	**Framework Version 2**	**#**	**Framework Version 3**	**Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/** **Focus Interview (FI)/ Survey data analysis (SDA)**	**Actions made to address the comments and results of the analyses**
FUNCTIONALITIES	25	Which of the following functionalities does the app make use of in order to provide its features? • The app supports personalised automated coaching or goal setting. • Use of in-built basic phone functions for users e.g. to calculate dosages, to set reminders, schedule events, make to do list etc. • Use of in-built advanced phone functions (e.g location detection using GPS, camera, phone's scanner, phone's note taking or voice recorder, Bluetooth for device connectivity, calculator) • Ability to transfer data (import and / or export) as required on request (and if relevant to the app) 1= Cannot be assessed or met none of the criteria above.	23	Which of the following functionalities does the app make use of in order to provide its features? •The app supports personalised automated coaching or goal setting. •Use of in-built basic phone functions for users e.g. to calculate dosages, to set reminders, schedule events, make to do list etc. •Use of in-built advanced phone functions (e.g location detection using GPS, camera, phone's scanner, phone's note taking or voice recorder, bluetooth for device connectivity, calculator, microphone and speaker) •Ability to download/transfer data (import and / or export) as required on request (and if relevant to the app)	**FB**: '-should add "microphone and speaker" to the list in point 3 Scoring for apps should be different depending on the level of its functionality (e.g., peer app/communication app vs. mood tracker app and etc.) **IRR:** Moderate agreement For top-rated app (ICC, Alpha >0.7), the level of reliability was lower *User Experience* A further refinement might be required for questions 9-12 and 24-27, which will be confirmed in approach 3 (analysis for the consistency within domains). A further refinement might be required for questions 9-12 and 24-27. **FI: 9.** *User Experience* cannot be assessed by evaluators. We recommend not to evaluate this domain in the framework or rather undertake a pilot exercise/review of the authenticity of end-user opinions or verbatim feedback and app	Added microphone and speaker in point 3

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met four of the criteria above.		1= Cannot be assessed or met none of the criteria above. 2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met four of the criteria above.	recommendations. However, a literature search or a guideline should be provided on how to check the authenticity of those reviews.	
EASE OF USE	26	Is the app easy to use? • Help page is available if there are any questions related to using the app. • The app provides overlay guidance or instructions on how to use the app (e.g. automatically help you navigate when you first use the app). • The app's search functions allow users to easily find what they are looking for. • The app has an intuitive design and users do not need further assistance in navigating the app.		Deleted	**FB:** '-Option 3 is unfair if the App does not need a search function, this limits the scoring to a Max of three criteria. -Ease of Use - if the app is simple but doesn't support Search (because it doesn't need it), it still only scores a 3. -if it is intuitive and easy to use without further instruction, it should be scored at the highest level rather than just as one of the criteria -it should say “Help or Tips” rather than just “Help” **IRR**: Slight Agreement	

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		1= Met none of the criteria above. 2= Met one of the criteria above. 3= Met two of the criteria above. 4= Met three of the criteria above. 5= Met all of the criteria above.			For top-rated app (ICC, Alpha >0.7), the level of reliability was lower *User Experience* For the items' reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24 and 26 are the question with poorer reliability within the domains. A further refinement might be required for questions 9-12 and 24-27, which will be confirmed in approach 3 (analysis for the consistency within domains). A further refinement might be required for questions 9-12 and 24-27.	
PERSONALISATION/ CUSTOMISATION	27	To what extent can the app be personalised? (A) App cannot be assessed (B) The app does not allow customisation/personalisation. (C) Basic personalisation is provided such as email reminders; avatar notification settings.	24	To what extent can the app be personalised? 1= App cannot be assessed 3= The app does not allow customisation/personalisation. 5= The app allows customisation/personalisation of some or all features	**FB:** it had the ability to personalise which modules you could download. However, there weren't the actual features of the app, such as notifications/reminders/ avatars etc. if the app interface is very simple, then this is not relevant Met D but not C - which option to select is not clear based on parentheses	The wording was made simpler

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
		(D) Additional personalisation of language choice, information, goal setting is available. (E) Personalisation of all the aspects of the app is available. 1= Cannot be assessed 2= The app does not allow customisation/personalisation 3= Met criteria (C) only. 4= Met 2 of the above criteria (C and D only) 5= Met 3 of the above criteria (C, D, E)			Personalisation - app allows you to favorite hypnosis recordings. I wasn't sure where this fit in but thought this was a basic function, so picked C. **IRR**: Fair agreement For top-rated app (ICC, Alpha >0.7), the level of reliability was lower *User Experience* A further refinement might be required for questions 9-12 and 24-27, which will be confirmed in approach 3 (analysis for the consistency within domains). A further refinement might be required for questions 9-12 and 24-27.	
Other						
					FB: App (Prepare for Stress) was free, but developer pushes a training program - there was no way to specify this. Also, it is intended to share data with researchers if enabled - there was no option to select for that use of data	

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
					The ability to import/export is in multiple questions Clustering the questions into what you should assess prior to download - what you should assess during download (e.g. during registration, did it ask if you were 18) and then what to assess within the app would be very useful. The evaluation took a very long moving between the app and the App Store all the time. There should also be an option to tick "unsure" for the questions. If there are some areas where you are not able to find the correct information or are not sure what you are looking at, then I wonder whether the scores will be accurate. An option of "unsure" allows this section to be reviewed by a second person for clarity. this app (PsychologyCompass) was mainly for mental performance, which included stress as part of it, and this app could fall into general mental health + anxiety	

Domain Sub-domain	#	Framework Version 2	#	Framework Version 3	Feedback from the evaluators (FB)/ Inter-rater reliability analysis (IRR)/ Focus Interview (FI)/ Survey data analysis (SDA)	Actions made to address the comments and results of the analyses
					I prefer questions to be on one page, so when I find the information in the app, I can flick down t the question and fill it in. In this format, you spend a lot of time going backward and forwards. I have mentioned this in another questionnaire, but possibly having an "unsure" option would be helpful, or "unable to locate" and some apps were vague. In the scoring, being able to see what criteria were selected rather than 'if 2 options or 3 options' for example, if it is important to see based on what criteria, the app was rated that score for each question.	

SCREENING

FRAMEWORK Version 2	Framework Version 3	Feedback	Actions made
Please provide the smartphone platform, model and operation software you have used to do the evaluation (i.e Android-Samsung, 8.5 or Apple iOs-iPhone 10, 14.4):	Not yet revised	1. Should include if the developer weblink is working. 2. Check Apple's data privacy feature. 3. App website is now deregistered. 4. Would be useful to have that as a screening question so any further evaluation should cease. 5. Developer website was not secure; privacy statement or T&C was not available. 6. What if privacy statement is not in English/ Privacy statement was not in English, so privacy related questions were hard to assess. (Need to check the Language of these statements in screening) 7. Some apps require birthdate and gender. We recommend for evaluators to use fictional DOB and random gender 8. When categorising, need to say what is the main subcat and not the accessory subcat need generic email for registration we should also note the category of the App e.g., if it is Medical, Health, Entertainment etc	Not yet revised
A) PRE-SCREENING QUESTIONS (exclude the app if either one of the criteria below are not met) Yes=1 OR No=0 1. App is related to the health area of interest 2. App supports English language 3. App is usable in Australia (i.e. download and accessible without the requirement for user to enter zipcode, phone number from another country 4. App is available without prescription by health professionals (such as app requiring access codes or password) 5. App updates have been made available within the last 18 months			
B) SCREENING QUESTIONS 6. Does the App load immediately without obvious crashing or freezing or other poor user interface experience?			

Decision: Proceed/need for a second device or second reviewer 7. [Second Review] Does the App load immediately without obvious crashing or freezing or other poor user interface experience? Decision: Proceed Evaluation/Stop Evaluation		9. Also we should make a decision on whether to screen out USA apps 10.Excluding apps without secure developer website and availability of T&C or privacy statement 11. When an app is broken in parts - brings up 404s etc. It would be good to be able to stop the assessment and say not recommend. 12. For an app, the developer contact has now left Flinders University. The email I sent to test has bounced back. I think this disqualifies the app (add as a screening q) Moved from evaluation to screening * The name or organisation and contact details of the app developer is available. * The name or organisation and contact details of the app developer is NOT available within the app or its supporting information. (e.g. e-mail address, phone number, address, or developers website)	

Appendix 2: Summary of the focus group analysis

Themes	Sub-themes	Key Findings
1. The need for health apps governance	-	• More regulations on health apps are needed in Australia due to uncertainty as to whether health apps in the Australian market are based on evidence or potentially carry information that may result in harm
2. General points for framework	2.1 Audience of the framework	• Priority of domains would be different with micro and macro-level perspective
	2.2 Opposing views related to importance of domains - The results of the Poll -The results of the focus groups discussion	• *Privacy/Security/Ethics/Legal* and content validity was the most important domains • *Privacy/Security/Ethics/Legal* is a priority and a role of organisations/app maker. • Content validity is deemed critical for health apps. • From both user and organisational perspectives, *Value* was noted as highly important. • All or some domains are equally important.
3. Practical approaches on evaluating domains based on the focus group discussion	3.1 The real definition and importance of *Accessibility*	• *Accessibility* means not only language but also access for vulnerable people and people with difficulties
	3.2 Valid source of information	• Gold standard way of evaluating information validity is Randomised Controlled trials. • Content must be evidence-based (I.e., based on clinical guidelines, expert knowledge)
	2.3 Future possibilities	• Evaluation studies on the clinical effectiveness of health apps
	3.3 *Developer Credibility* - Who is the main driver of the health apps - Reputation and health app affiliation - Opposing views in evaluating Developers Credibility - Transparency	• Defining the 'Developer' is important in the framework. • The reputation of the app maker (history) is a criterion. • Affiliation with organisations is an indicator. • Apps should be peer-reviewed or confirmed by a health professional. • The credibility of developer does not mean a quality app. • Health apps users feel more secure with a credible organisation. • Transparency of app maker is *Value*d (i.e., Publishing information online)
	3.4 *User Engagement* - Gamification - *User Engagement* and content personalisation - Measures of *User Engagement*	• Gamification is one of the elements which makes an app distinctive. • There are concerns over addictive behaviour due to gamification. • The audience for the content and its relevancy to them is important for the *User Engagement* of an app. • The real use of the app and engagement can be assessed by conducting a RCT to get user ratings or looking at previous studies.

		• Other measures would be the number of people who downloaded and logged in or registered, how often a user access an app, and how long they stay on the app. • The frequency of activities done by users within the lifecycle of an app is also important. • Customer verbatim reviews on health apps is a good indicator.
	3.5 *Privacy/Security/Ethics/Legal* - Complexities in evaluating *Privacy/Security/Ethics/Legal*ities - User capacity to review the guidelines - The real responsibility is on the organisation (app driver) - Transparency - An area for isolated regulation	• Likelihood of reading a privacy statement is very low, and it would be beyond customer capacity. • The responsibility lies on the organisation/app maker, not the app user. • Ethics is the most important from an organisational perspective. • App users expect all privacy/security/ethical and legal matters are already addressed by app driver. • App driver should ensure complying with national standards and regulations • Transparency is important in storing data. • The area around the regulation of apps in Australia needs to be improved.
	3.6 *User Experience* -	• *User Experience* is hard to measure by another person on behalf of the app user. • It should be measured using app users' opinions. • User recommendations of apps to others is a good indicator of *User Experience*.
	3.7 *Value* -	• *Value* domain should be clearly defined as it differs based on the perspective. • From an individual level, *Value* means the perceived benefits of the apps. • *Value* can be related to the clinical outcomes resulting from using the app. • *Value* domain could not be simply measured by its monetary *Value*. • Some users perceived purchased apps as more valuable than free apps. • It is difficult to measure *Value* domain, and it needs some validation study. • Some suggested evaluating the usability of the app or whether it addresses the problem of the user.
4.Relationships between domains	4.1 The link of *User Engagement* and *User Experience* to *Value*	• *User Experience* is *Value*-driven from the individual perspective. • Gamification comes down to the *Value* of the app, which is what makes the app valuable and realistic for the user.
	4.2 The link of *Accessibility* to *User Experience* and *Value*	• *Accessibility* and usefulness can be used to measure *User Experience*. • *Accessibility* was also mentioned as an indicator to assess *Value*.
	4.3 *The link between the Clarity of Purpose of the App from organisational perspective and Value*	• The scope (intended purpose) of the client organisation is a factor when choosing a particular app for a curated set of apps. (i.e., customer attraction or benefits to the users).
	4.4 The link of all domains to *Value*	• Assessing other domains for an app would ultimately mean an assessment of *Value* of the app.

Appendix 3: Mental health apps for evaluation

Appendix 3.1: The number of apps derived from screening

	Depression	**Anxiety and stress**	**General mental health**	**Others ***	*Depression, anxiety, General mental health*	*Depression, anxiety, other*	*Depression, anxiety, General mental health, other*	**Total**
Depression	**4**	**9**	**3**					16
Anxiety and Stress		**36**	**45**	**1**				81
General Mental health			**113**	**1**				114
Others*				**32**				32
Depression, anxiety, General health					16			16
Depression, anxiety, other						1		1
Depression, anxiety, General mental health, other							1	1
Total	4	45	161	34	16	1	1	261

Appendix 3.2: The list of 49 Apps included in the Evaluation

No.	**App Name**	**Developer**
1	Self-manage Depression: Daily exercise (GGDE)	Ggtude LTD
2	feel better - Mood & CBT therapy to manifest goals (Andoid) Fee better-Moods & goals (Apple)	Media Local Studios
3	WorryTree: Anxiety Relief & CBT Diary	WorryTree Ltd
4	WellMindSA	Runninghill Software Development PTY LTD
5	Unwinding Anxiety®	MindSciences, Inc.
6	7 cups: Online therapy & Chat	7 Cups of Tea
7	Stress Control Norbu	Aleksandr Zheleznov
8	Panic Attacks or Anxiety?	Geert Verschaeve
9	Tappily	Dalton Technologies, Inc.
10	Anxiety Reliever: Mental Health Support	Madlen Fedorova
11	Serenita - Stress & Anxiety	Eco-Fusion

12	Android: Resility Stress Management & Biofeedback Apple: Resility Personal Biofeedback	Resility Health LLC
13	Prepare for Stress	Sara Denning
14	PanicShield - Panic Attack Aid	Inquiry Health LLC
15	Ease My Stress & Anxiety	Disease & Stress Free Lifestyle CluB LLC
16	idstress - Overcome stress (Android) idstress-meditate, find calm	Ana Lombard
17	MindSurf-Manage Stress	Tim Carey
18	Wim Hof Method	Innerfire B.V.
19	PsychologyCompass	PsychologyCompass
20	Primed Mind - Motivation, Mindset & Life Coach	Primed Mind GmbH
21	Rootd - Panic Attack & Anxiety Relief (Android) Rootd - Panic Attack Relief (Apple)	Simply Rooted Media Inc
22	Thinkladder - Self-awareness & Mental Wellness	Think Ladder Limited
23	Clear Fear	Stem4
24	Dare: Anxiety & Panic Attack Relief	BMD Publishing Ltd
25	My Safe Zone - Anxiety Attack Assistance	Rock IT 88
26	Synctuition Mindspa Meditation	Synctuition OÜ
27	SwipeJoy: Self Growth and Mental Wellness with Fun	SwipeJoy
28	wingwave	Besser-Siegmund-Institut GmbH
29	S.T.A.R. Support	OptimumU Health and Wellness Inc
30	Address Stress	Compact Health PTY LTD
31	Pocketcoach - Reduce Stress (Android) Pocketcoach-Anxiety Helper (Apple)	Pocketcoach GmbH
32	Fear Buster: Deep Relaxation and Stress Relief	Julie mccraken
33	WorryTime by ReachOut	ReachOut Australia
34	Be Okay	Helena Leitao
35	The Stress Code	Fusebox
36	UrbanYogi	Reverberation Tech Private Limited
37	Woebot	Woebot Labs Inc
38	Fuzing: Relax, De-Stress & Sleep Well - Free (Android) Relax, De-Stress & Sleep Well (Apple)	Fuzing LLC
39	Lift – Depression & Anxiety	Impact Collective, LLC
40	Youper AI Therapy for Anxiety & Depression (Android) Youper: Self-Guided therapy (Apple)	Youper, Inc.
41	UpNow Hypnosis	Renewed Edge
42	Smiling Mind	Smiling Mind
43	Mind Ease: Anxiety Relief	Mind Ease Labs LTD
44	Shine: Calm Anxiety & Stress	Shine

45	Breath Ball: The Stress Relieve Breathing Exercise (Android) breath ball Breathing Exercise (Apple)	Michael Holl
46	MindFull: Anxiety Tracker	Brady Simon
47	Headspace: Meditation & Sleep	Headspace inc
48	MindShift CBT - Anxiety and Panic Relief (Android) MindShift CBT - Anxiety Relief (Apple)	Anxiety Canada Association
49	InnerHour Self-Care Therapy: Anxiety & Depression	InnerHour

Appendix 3.3: The top 10 apps based on D-HEAL FMHE

No	App name in Apple store (Developer)	App name in Google Play (Developer)	Average score	Disease Category**	1: Has information with topic area, 2: Meditation/ mindfulness/Mood Tracker, 3: Both	Brief description of the app
1	Clear Fear (Stem4)	Clear Fear (Stem4)	97.33	Anxiety and Stress	3: Both	The app noted that it could be used by 11-18 years old and adults to help with anxiety and stress through CBT incorporated activities, anxiety tracker, goal setting, and resources. However, we recommend that minors use the app with the supervision or support of a parent or carer.
2	Headspace: Meditation & Sleep (Headspace Inc.)	Headspace: Meditation & Sleep (Headspace Inc.)	94.33	Anxiety and or stress, general mental health	2: Meditation/mindfulness/Mood Tracker	The app is for adults and is used to help with stress anxiety and boost general health through guided meditation and mindfulness exercises. Individuals under the age of 18 may utilize the app only with the involvement and consent of a parent or legal guardian.
3	Smiling Mind (Smiling Mind)	Smiling Mind (Smiling Mind)	92.33	Anxiety and or stress, general mental health	3: Both	Free mindfulness meditation app used to assist people dealing with pressure, stress and challenges of daily life. The app can be used for people ages three and beyond.
4	Unwinding Anxiety (MindSciences Inc)	Unwinding Anxiety (MindSciences Inc)	88	Anxiety and Stress	3: Both	The app is for adults experiencing anxiety. Users can share experiences, access videos and lessons, and use the app for journaling. Users between 13 to 18 are recommended to ask their parent/guardian permission before using the app.

5*	MindShift CBT - Anxiety Relief (Anxiety Canada Association)	MindShift CBT - Anxiety and Panic Relief (Anxiety Canada Association)	87.67	Anxiety and or stress, general mental health	3: Both	A self-help anxiety relief app for adults based on CBT. The app provides anxiety information and allows users to practice CBT therapy. For minors, the app should only be used if a healthcare professional has suggested it AND that the parent/guardian is aware of the app use. *Caution: PROVISIONALLY RECOMMENDED. Please note that the Apple store stated that the developer had not provided privacy practice and handling data to Apple; this information will be required in the next update. However, a privacy statement was available via the developer's web link and within the app.*
6*	Rootd - Panic Attack Relief (Simply Rooted Media Inc.)	Rootd - Panic Attack & Anxiety Relief (Simply Rooted Media)	87.67	Anxiety and Stress	3: Both	The app is for adults to help in overcoming stress, panic attacks through a guided deep breathing exercise, meditation, journaling, and visualisation. Individuals under the age of 18 may utilize the app only with the consent of a parent or legal guardian.
7*	Norbu Stress Control & Breathe (Aleksandr Zhelenzno)	Norbu: Antistress training, breathe, meditation (Mindfulness App Dev)	87.67	Anxiety and Stress	2: Meditation/mindfulness/Mood Tracker	The app is for users over 13 years old. It has mindfulness-based stress control techniques that help manage stress, including anxiety relief games, breathing exercises and guided meditation.
8	Mind Ease: Anxiety Relief (Mind Ease Labs Ltd)	Mind Ease: Anxiety Relief (Mind Ease Labs Ltd)	82	Anxiety and or stress, general mental health	2: Meditation/mindfulness/Mood Tracker	An anxiety relief self-care app for adults (16 years and below should not use the app without a parent or guardian) that offers different exercises such as mindfulness meditation, deep breathing visualisation, anxiety defusion. It also offers inspirational quotes and allows users to track their progress.

9	WorryTime by ReachOut (ReachOut Australia)	ReachOut WorryTime (ReachOut Australia)	81.33	Anxiety and Stress	2: Meditation/mindfulness/Mood Tracker	The app is to help anxiety, stress and worries by allowing to set times to release worries based on CBT. The age recommendation for use was not reported. *Caution: PROVISIONALLY RECOMMENDED. Please note that the Apple store stated that the developer had not provided privacy practice and handling data to Apple; this information will be required in the next update. However, a privacy statement was available via the developer's weblink.*
10	Pocketcoach-anxiety helper (Pocketcoach GmbH)	Pocketcoach - Reduce Anxiety (Pocketcoach)	81	Anxiety and Stress	3: Both	Self-help for adults (at least 18 years old) experiencing anxiety, stress or panic attacks based on CBT, mindfulness and ACT. *Caution: Please check the privacy statement link within the app. The link attached in the app description is currently not working; however, Apple Store has stated that the app has provided information about privacy practices and data handling.*

Appendix 3.4: Average scoring for the 49 mental health apps

App Name	Average score
Clear Fear	97.33
Headspace: Meditation & Sleep	94.33
Smiling Mind	92.33
Woebot- Your Self-Care Expert	89.33
Unwinding Anxiety	88.00
MindShift CBT - Anxiety Relief	87.67
Rootd - Panic Attack Relief	87.67
Stress Control Norbu	87.67
breath ball Breathing Exercise	86.00
Youper: Self-Guided therapy	85.67
Lift – Depression & Anxiety	85.00
Mind Ease: Anxiety Relief	82.00
WorryTime by ReachOut	81.33
Pocketcoach: anxiety helper	81.00
WorryTree: Anxiety Journal	80.67
Wim Hof Method	80.33
Synctuition Mindspa Meditation	79.00
Dare: Anxiety & Panic Attack Relief	77.67
S.T.A.R. Support	77.33
7 cups - Anxiety & Stress Chat	76.33
Thinkladder	76.00
Serenita - Stress & Anxiety	75.33
Shine: Calm Anxiety & Stress	75.00
UrbanYogi	75.00
InnerHour Self-Care Therapy: Anxiety & Depression	74.67
PsychologyCompass	74.67
wingwave	73.67
Address Stress	72.33
SwipeJoy: Self Growth with Fun	72.33
Self-manage Depression-GGDE	72.00
Primed Mind - Mindset Coach	70.67
UpNow Hypnosis	70.00
Resility Personal Biofeedback	69.33
Tappily	68.67
MindSurf-Manage Stress	67.33
My Safe Zone - Anxiety Attack Assistance	67.00
PanicShield - Panic Attack Aid	67.00
Relax, De-Stress & Sleep Well	65.67
Anxiety Reliever: Mental Health Support	65.00
MindFull: Anxiety Tracker	65.00
Be Okay	63.33

Fee better-Moods & goals	63.00
Panic Attacks or Anxiety?	62.33
Fear Buster: Deep Relaxation and Stress Relief	62.00
Prepare for Stress	60.00
idstress-meditate, find calm	58.33
The Stress Code	55.00
Ease My Stress & Anxiety	54.33
WellMindSA	53.33

Appendix 4: ICC Analysis

Subgroup analysis

Approach 1: Per rated apps

Among 49 apps rated by three raters using 27 questions for ten domains, ICC analysis for individual apps indicated a statistic of up to 61% of rated apps with a high level of consistency in rating (ICC, Alpha > 0.7, n= 30). An alpha of >0.7 was observed in all items and overall rating in these apps. When items attributed to odd single Alpha scores were removed, overall average Alpha scores still remained as 0.7 in all apps. The graph below showed the distribution of the 30 apps, which displayed a "high" reliability score with a 95% confidence interval (Figure 5). Three other apps showed "good" agreement between raters (ICC >07); however, the consistency between each item within these apps was low. ICC analysis for each app indicated some individual items whose removal would lead to a higher/lower score in the overall ICC rating. For example, Q2 was suggested to be removed from the rating of app ID 1047 for a higher Alpha score in this app (0.825 vs 0.733), while Q6 was suggested to remove for a poorer ICC overall score (0.589 vs 0.733). The remaining 16/49 apps have either Alpha <0.7 or negative alpha, indicating poor consistency between agreements. In these apps, a high volume of items was suggested to be removed for a higher overall alpha score in each app; therefore, further investigation in approach 2- analysis per item and analysis 3- per items and domains is required.

Approach 2: Per items

The 16 apps with "poor" ICC score (ICC< .7) and low consistency for each rated item were selected for the second analysis. Analysis stratified by questions showed the distribution of poorly rated questions in the poorest rated apps (Figure 6). Question 1-3, 9-15, 17, 24 and 26 had the highest frequency (>20%) and appeared to be problematic questions in poorly rated apps (original total Alpha score of the app <0.7, and total Alpha change to over 0.7 if the item is removed), comparing to items that have been removed due to the high consistency without covariance (e.g., all raters have same rating score). Question 17 frequently displayed "poor" agreement between raters, followed by question 11 and question 14 (50% and 31%, respectively). Question 23 was the best consistency amongst raters even in poorly rated apps, followed by Q4, 5, 8, 16 and 22. The remaining questions fell within the standard deviation expected and acceptable for this data set.

Sub-group analysis for the poorest rated apps (i.e., ICC<0.7, high prevalence of alpha score <0.7 if items are removed) indicated an over 55% times appearance of questions 1-3, 6, 7, 9-12, 13-15, 17, 24-26 (Q9 and Q17= 100%) in these poorest rating apps. This suggests that further refinement is needed for these questions, and thorough training for participated raters is required to overcome the issue.

The analysis also showed a large negative Alpha *Value* reported in all questions (Q1-27, n= 9 apps, 3 raters for each app, frequency = 11.11%-55.56%), suggesting a negative average covariance among items.[3] This violates reliability ICC model assumptions; therefore, attention also needs to be paid to data quality. A possible explanation was due to the sampling error that was attributed by the small number of raters available for each app in this Framework evaluation exercise (n=3)

Approach 3: Per domain and item with domain

The analysis indicated that the *Privacy/Security/Ethics/Legal* domain has the poorest average ICC result in poorly rated apps (n=10/14), followed by the *Accessibility* domain (n= 9/14). Content Validity, Technical aspects and features, and *User Experience* equally have 6/14 times that the domain has poor internal consistency (n=6/14). A similar observation was not seen in *Clarity of the Purpose of the App, Interoperability, User Engagement*, and *Value* due to limited items available for analysis (one to two questions). For the top-rated app (ICC, Alpha >0.7), the level of reliability was lower in domain *Developer Credibility*, *Accessibility*, and *User Experience*.

In consideration for items within each domain, all items within *Privacy/Security/Ethics/Legal* appeared to have unreliable items showing a poor agreement for this domain (removing items will increase the alpha score: Alpha toward 0.7). In contrast, the opposite direction was observed in *Content/Information Validity* (removing items will lead to a decrement in the alpha score, Alpha moved toward .7). These items of *Privacy/Security/Ethics/Legal* have the highest frequency of being removed for a higher average Alpha score in poorly rated apps. Interestingly, on some occasions, removing the item(s) within this domain leads to a decrement in the average Alpha score below 0.7; thus, the level of agreement in this domain is inconclusive. Extra caution needs to be paid in interpreting the results of this analysis approach due to the following reasons:

For all rated apps, only two domains had sufficient statistical evidence (*Content/Information Validity* and *Privacy/Security/Ethics/Legal*) to allow interpretation and reporting. Statistical evidence in other domains was not reliable and meaningful due to the inability to conduct a statistical test due to insufficient sample

Among the items rated within these domains (*Content/Information Validity* and *Privacy/Security/Ethics/Legal*), a high volume of negative scores was also observed, indicating an insufficient sample size for the test (sampling error produced negative average covariance that due to small sample size associated with the number of rater for each app in this Framework evaluation (n=3)

On the other hand, lower frequency or absence of similar issues was observed in all other domains, suggesting that *Privacy/Security/Ethics/Legal* might be seen as an independent domain within the Framework for more rigour in the overall rating.

The results of ICC and Fleiss's Kappa are quite consistent. Absolute agreement in items was detected for question 23 in both analyses. However, inconsistency between analyses was observed in some apps. For example, apps 1135 and 1374 had a moderate agreement in Fleiss's Kappa analysis but were poorly rated in ICC analysis. As Indicated in ICC analysis 2, these apps have an ICC negative score that was potentially attributed to being rated by a smaller number of observations for each app; thus, consideration is required upon interpretation.

For the items' reliability, both analyses indicated Questions 1-3, 9-13, 15, 17, 24 and 26 were the question with poorer reliability within the domains. The results on only one question (Q14) were not consistent across the two approaches of analyses (Fleiss's Kappa vs ICC). In contrast, Fleiss's Kappa indicated a substantial agreement for Q14. However, in the ICC approach, Q14 was seen as having a high agreement in top-rated apps, but poor agreement in the low rated app, indicating poor agreement in the poorest rated app can be solved by detailed training. Similarly, although Q2-3, 13,15,17 seems problematic, the multi-approach analysis showed that these issues occur more prevalent in the poorer rated app; thus, detailed training can be considered to resolve the issue.

The analysis indicated a high consistency in rating outcomes for our framework. Pool observation indicated a high level of consistency between observation with ICC= was .855, 95% CI (.819-.886). Considering each item (question) within the framework, most items showed high consistency and were less problematic. If these items are removed, the Alpha score of the framework will decrease. However, pool analysis detected four questions that potentially can be improved upon further investigation (Q8,16-17 and 20).

Furthermore, 61% of apps rated using the framework had an ICC score of over 0.7 (good reliability) in the whole Framework analysis approach with consideration to items suggested to be removed to improve consistency score (Alpha >0.7). Among poorly rated apps, a high volume of items removed to archived Alpha score over 0.7 and a negative score for average/ individual items was seen. This indicated a sampling error for poorly rated apps that led to negative average covariance amongst raters. The issue can be resolved by increasing observation for each app in the future framework refinement test. Literature suggested observation of between 7 and 10 raters for any three-four random apps (n=3-4), using D-HEAL 27 items (questions) framework can be sufficient to address sampling error (80% power to detect ICC=$0.7^{)4}$. Subgroup analysis in each item and domain showed the opportunity for improvement in questions 1-3, 9-13, 15, 17, 24 and 26. for the whole framework approach. However, the issue of inconsistency can be resolved by offering detailed training for questions Q2-3, 13-15 and 17. A further refinement might be required for questions 9-12 and 24-27, which will be confirmed in approach 3 (analysis for the consistency within domains)

Approach 3 analysis indicated that *Privacy/Security/Ethics/Legal* consists of questions 9-12) has the poorer average ICC result in poorly rated apps (n=7/10) and good rated apps, suggesting improvement is required for this domain. Due to the equivocal result for *Privacy/Security/Ethics/Legal* in this exercise for this framework, an option to consider this domain as an independent domain is available due to the acceptable/good consistency in other domains.

Appendix 5: Demographic information about the survey responders

	Baseline (S1) (n=55)		**Follow-up 1 (S2) (n=27)**		**Follow-up 1 (S3) (n=23)**	
	No	%	No	%	No	%
Gender						
Male	6	10.91	1	3.70	1	4.35
Female	48	87.27	26	96.3	22	95.65
Non-binary/third gender	1	1.82	-	-	-	-
Age						
18-25 years	7	12.73	4	14.81	3	13.04
26-35 years	15	27.27	7	25.93	6	26.09
36-45 years	9	16.36	2	7.41	2	8.70
46-55 years	13	23.64	8	29.63	6	26.09
56-65 years	9	16.36	5	18.52	5	21.74
66-75 years	2	3.64	1	3.70	1	4.35
Relationship						
Married or Defacto	25	45.45	8	29.63	7	30.43
Never married	19	34.55	10	37.04	9	39.13
Widowed	2	3.64	2	7,41	2	8.70
Divorced	2	3.64	2	7.41	2	8.70
Separated	2	3.64	4	14.81	3	13.04
Prefer not to say	4	7.27	-	-	-	-
Did not answer (missing *Value*)	1	1.82	1	3.70	-	-
Education						
Postgraduate Degree	22	40.00	10	37.04	8	34.78
Graduate Diploma/Graduate Certificate	7	12.73	4	14.81	4	17.39
Bachelor Degree	16	29.09	8	29.63	7	30.43
Certificate	2	3.64	2	7.41	2	8.70
High School	8	14.55	3	11.11	2	8.70
Never attended school	0	0	-	-	-	-
Prefer not to say	0	0	-	-	-	-
Employment						
Full-time	23	41.82	10	37.04	8	34.78
Part-time	9	16.36	4	14.81	3	13.04
Self-employed	3	5.45	3	11.11	2	8.70
Student or training	5	9.09	2	7.41	2	8.70
Unemployed looking for work	5	9.09	2	7.41	2	8.70
Retired	6	10.91	2	7.41	2	8.70
Not working due to illness	4	7.27	4	14.81	4	17.39
Prefer not to answer	0	0	-	-	-	-
Income			-	-	-	-
Negative or zero Income	1	1.82	1	3.70	1	4.35

$1 - $9,999 per year	1	1.82	1	3.70	1	4.35
$10,000 - $19,999 per year	1	1.82	1	3.70	1	4.35
$20,000 - $29,999 per year	2	3.64	1	3.70	1	4.35
$30,000 - $39,999 per year	4	7.27	2	7.41	1	4.35
$40,000 - $49,999 per year	2	3.64	2	7.41	2	8.70
$50,000 - $59,999 per year	4	7.27	3	11.11	2	8.70
$60,000 - $79,999 per year	6	10.91	2	7.41	1	4.35
$80,000 - $99,999 per year	9	16.36	4	14.81	4	17.39
$100,000 - $124,999 per year	4	7.27	1	3.70	1	4.35
$125,000 - $149,999 per year	2	3.64	4	14.81	4	17.39
$150,000 - $199,999 per year	9	16.36	2	7.41	1	4.35
$200,000 or more per year	4	7.27	2	7.41	2	8.70
Prefer not to answer	4	7.27	1	3.70	1	4.35
Don’t know	1	1.82	-	-	-	-
Did not answer (missing *Value*)	1	1.82	-	-	-	-