Alberta Rating Index for Apps (ARIA): An Index to Rate the Quality of Mobile Health Applications

by

Peyman Azad Khaneghah

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in
REHABILITATION SCIENCE

Faculty of Rehabilitation Medicine
University of Alberta

Abstract

Introduction: The number of mobile health applications (m-health apps) available to the public through online application (app) stores is rapidly increasing. In addition, the general public's interest to use m-health apps as an adjunct to conventional health care services is increasing. However, because of the inadequate quality control mechanisms on app stores and lack of stringent regulations for m-health apps, there is a risk of apps to have inferior quality or even be harmful. In the absence of formal guidelines, users may choose apps based on unreliable information such as reviews and ratings on the app download page, or the number of downloads for an app. Rating scales to evaluate apps exist for m-health app users, health care providers, and researchers. Most are long and complicated scales that are not appropriate for the general public. None have been developed using a theoretical framework. The purpose of this thesis was to develop the Alberta Rating Index for Apps (ARIA) based on the theories of technology acceptance, frameworks of app evaluation, and lived experience of users of mobile health applications.

Methods: A multi strategy study was conducted in three phases. In phase one, the investigator conducted six focus groups with users of mobile health applications including older adults, adults with a mental health condition, health care providers, and app developers to identify quality criteria that were important to users and developers of m-health apps. Next, an item pool was generated based on a review of app rating scales. In phase two, the content of the item pool was validated using an online survey and a calculation of the content validation index for each item. Also in phase two, a sample of participants from the online survey participated in a focus group to shortlist the item pool and develop the first draft of ARIA. In phase three, ARIA was piloted by nine potential users of m-health apps, including older adults and adults with a mental health

condition. Also, in phase three, the inter-rater reliability and criterion-related validity of ARIA were examined using 16 participants consisting of 4 older adults, 4 adults with a mental health condition, and 4 health care providers. The scores of ARIA were correlated with the scores of users' version of Mobile Apps Rating Scale (U-MARS) to examine the criterion-related validity.

Results: Nine quality criteria measure the quality of m-health apps: the purpose of the app, trustworthiness, privacy, security, affordability, ease of use, functionality, appropriateness to target users, and usefulness and satisfaction. Generalizability coefficients (G-coefficients) were calculated using ARIA total scores as the measure of reliability. High G-coefficients for health care providers (G = 0.98), older adults (G = 0.83) and adults with mental health conditions (G = 0.88) indicated that users could reliably rate the quality of m-health apps based on total scores of ARIA. The positive but low correlation of ARIA's total scores with U-MARS indicated that both assessment tools measure the quality of m-health apps. However, the quality criteria were different between the ARIA and U-MARS. Participants in phase three reported that ARIA was easier and more convenient compared to U-MARS.

Conclusion: ARIA is the first mobile application rating index developed based on theories of technology acceptance and frameworks of app evaluation. ARIA was designed to be used by health care providers and the general public, including older adults, adults with mental health conditions, and family caregivers. The content of ARIA was validated through a rigorous process. Moreover, three types of app users tested the inter-rater reliability of ARIA. Users perceived ARIA to be easier and more convenient compared to U-MARS. Clinical implications of ARIA are to help patients, family caregivers, and healthcare providers rate the quality of mobile health applications and identify the ones that are acceptable. Health informatics

researchers may use ARIA to develop health apps that are useful and acceptable to end users, especially older adults.

Preface

This thesis is an original work by Peyman Azad Khaneghah. Dr. Lili Liu supervised the conceptual design of the dissertation, writing process, and funding for the study. Peyman Azad Khaneghah designed the study concept and experimental methods, undertook subject recruitment, data collection, analysis, and writing. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name "A Rating Scale for Mental Health Mobile Applications for Older Adults", Pro00067856, February 9, 2017.

Parts of Chapter 1 of this thesis has been published in the form of a systematic review of literature as Azad-Khaneghah, P., Neubauer, N., Miguel Cruz, A., & Liu, L. (2020). Mobile health app usability and quality rating scales: a systematic review. *Disability and Rehabilitation: Assistive Technology*, 1-10. doi:10.1080/17483107.2019.1701103. I was responsible for the data collection and analysis as well as the manuscript composition. Neubauer, N. and Miguel Cruz, A. assisted with the review of the literature as the second and third reviewers, and contributed to manuscript edits. Liu, L was the supervisory author and was involved with concept formation and manuscript composition.

Acknowledgement

First, I would like to express my deepest and most sincere gratitude to my Ph.D. Supervisor, Dr. Lili Liu, for the continuous support of my Ph.D. study and academic achievements, for her patience, motivation, enthusiasm, and immense knowledge. I could not have imagined having a better supervisor and mentor for my Ph.D. Second, special thanks to Dr. Mary Roduta Roberts, my co-supervisor, whose insight and in-depth knowledge of research methods and health measurement scales gave me a new perspective in this field. I would also like to thank the members of my supervisory committee, Drs. Eleni Stroulia and Martin Ferguson-Pell for their time, advice, and feedback. I am grateful for all that I have learned from them.

I want to thank Drs. Christine Daum, Antonio Miguel-Cruz, Adriana Rios Rincon and Noelannah Neubauer, for their help and support in this project. I have been privileged to work with them in the same research team. I have never seen such a genuine camaraderie.

I would like to thank the Aging Gracefully across Environments using Technology to Support Wellness, Engagement, and Long Life Network of Centres of Excellence (AGE-WELL NCE), the Alberta Association on Gerontology, the Alberta College of Occupational Therapists, and the Faculty of Rehabilitation Medicine – University of Alberta, for funding to support this work.

My gratitude to the Organization for Bipolar and Affective Disorders in Calgary (OBAD), Seniors Association of the Greater Edmonton (Sage), Canadian Back Institute (CBI) clinic in Calgary, and Mental Health Commission of Canada, for their support. Special thanks to Mr. Kaj Korvela, Ms. Elizabeth Anderson, Ms. Anna Aishford and all participants in this project without whom this project could never happen.

Finally, I would like to thank my wife, Dr. Ailar Ramadi, for her patience, support, and love. I cannot thank her enough for all she has done for me.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AHS: Alberta Health Services

App: Application

CBI: Canadian Back Institute

CIHR: Canadian Institute of Health Research

COPD: Chronic Obstructive Pulmonary Disorders

CSUQ: Computer System Usability Questionnaire

DSM-5: Diagnostic and Statistical Manual of Mental Health Disorders - 5th Edition

FDA: The United States Food and Drug Administration

Health ITUES: Health Information Technology Usability Evaluation Scale

HIMSS: Health Care Information and Management Systems Society

iOS: Apple Inc. operating system

MARS: Mobile Applications Rating Scale

MDRs: Canadian Medical Device Regulations

MedAD-APPQ: Medical Adverse Effect App Quality Scale

MHCC: Mental Health Commission of Canada

M-Health app: Mobile health application

NHS: National Health Services (United Kingdom)

OBAD: Organization for Bipolar and Affective Disorders

ORCHA: Organization for the Review of Care and Health Applications

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta Analyses

PUEU: Perceived Usefulness and Ease of Use Questionnaire

QAEM: Quality Assessment Tool for Evaluating Medical Apps

SAGE: Seniors Association of Greater Edmonton

STAM: Seniors Technology Acceptance Model

SUMI: Software Usability Scale

SUS: System Usability Scale

TAM: Technology Acceptance Model

U-MARS: Users Version Mobile Applications Rating Scale

UTAUT: Unified Theory of Acceptance and Use of Technology

CHAPTER 1: Introduction and Review of Literature

1.1. Mobile Health Apps on App Stores

Mobile applications (apps) are discrete and independent software that run on mobile devices, such as smartphones and tablets [1]. Apps have similar capabilities to programs that run on personal computers and enable users to browse the internet, process documents, and play videogames on mobile devices such as smartphones or tablets [2].

Since 2008, users can download mobile apps from online markets known as app stores. Google Inc. and Apple Inc. dominate the app markets with their commercial app stores Google Play Store and iTunes, respectively. Apps on Google Play Store work on tablets and phones that run the Android open-source operating system. Apps on the iTunes work on Apple's proprietary operating system, the iOS. Google's Android operating system claims 82.8% of global market share while Apple's iOS owns 13.9% [3]. The remaining market is shared between BlackBerry, Nokia, and Windows phones. Both Google Play and iTunes provide various types of apps, including books, music, educational apps, games, and health and wellness apps [4].

A considerable number of apps on the app stores are related to health and wellness. The literature refers to them as mobile health apps, e-health apps, and wellness apps. For convenience, they will be referred to as "mobile health apps" or "m-health apps" in this dissertation. M-health apps can usually be found under Health and fitness, Lifestyle, and Medical categories on the app stores, in both iTunes and Google Play Store [5].

The number of m-health apps on app stores has increased rapidly over a period of three years. In 2014 there were more than 100,000 m-health apps worldwide [6]. This number increased to 325,000 in 2017 [7]. A report by the *Institute for Health Care Informatics* [8] estimates that almost 70% of m-health apps are relevant to general health and wellbeing such as

diet, lifestyle and stress management, and fitness while the rest are relevant to specific diseases, medication information, reminders, and women's health and pregnancy.

A considerable number of m-health apps on app stores are relevant to mental health. According to the *Institute of human data science* [8], mental health apps lead the number of apps under the disease-specific category. In another review, Donker and colleagues [9] reported that 24% of health apps on iTunes app store are relevant to mental health, including both disease-specific apps and apps on outcomes related to mental health (i.e., sleep, meditation, and mindfulness). According to the U.S. *National Institute of Mental Health* [10], the current trends in mental health apps include apps for self-management of symptoms, apps for improving thinking skills, apps for enhancing social skills, and apps used for passive tracking of a patient's symptoms.

1.2. Quality of Mobile Health Apps

The booming number of m-health apps has been reported as consumer-led and commercially driven rather than based on scientific evidence [11]. Many mobile health apps have not been validated, and the research for these apps lags behind their proliferation and adoption by the public [12]. A review of *National Health Services* (NHS) Apps Library in the United Kingdom showed that only 15% of the apps on the library provided evidence that they were effective [11]. Moreover, in a systematic review, Donker et al. [9] reported that the majority of m-health apps were not developed based on scientific evidence. Content of most of the mobile health apps also does not align with practice guidelines or established therapeutic principles [13].

Another concern is the extent to which m-health apps can protect the confidentiality of app users' information. Many apps do not have proper security measures such as data encryption

or password protection [12]. In the absence of security measures, users' information can be accessed by unauthorized individuals through digital theft or loss of a mobile device. Also, many m-health apps do not explain what information is collected and how the collected data is handled. In a review of apps for bipolar mood disorder, only 18 out of 82 apps had a privacy policy [13].

Finally, there is a risk for m-health apps to provide inaccurate or harmful information. In a review of online app stores, Bindhim et al. [14] identified several apps that promoted the use of illicit drugs. Such concerns create an unclear prospect about the quality and safety of m-health apps used by patients and health care professionals [15].

1.3. Older Adults and M-Health Apps

The challenges associated with identifying high-quality apps may impact older adults more than other cohorts of app users. In recent years the number of older adults who are using smartphones and tablets has increased [16, 17]. At the same time, older adults as a cohort may face a spectrum of health issues such as musculoskeletal, cardiac, metabolic, cognitive, and mental health which makes them an important target group for health apps. Moreover, older adults may face barriers when using smartphones and tablets due to age-related declines in sensory, motor, and cognitive skills. These limitations can affect the use of technology in this population. Apps that run on touch screen interfaces need to be adapted to the needs and limitations of these users. Issues such as small screen sizes of smartphones combined with touch screen navigation and complex interactions (e.g., pinch, swipe, or double tap) can be challenging for adults with age-related perceptual and sensory changes [18].

1.4. Regulations on M-Health Apps

Although the number of m-health apps is growing, limited regulations exist for m-health apps in industrialized countries. For example, the Food and Drug Administration (FDA) in the United States only regulates apps that are either "used as an accessory to an FDA regulated medical device" (e.g., apps that allow physicians see medical images on a tablet), or "transform a mobile platform into a regulated medical device" (e.g., apps that turn an iPhone to an electrocardiograph machine) [19]. In Canada, medical devices for human use are regulated by Health Canada according to *Medical Device Regulations* (MDRs), which were made under the *Canada Food and Drugs Act* (1985) [20]. The Canadian medical device regulations do not specifically name mobile apps as medical devices. However, a mobile app may qualify as a medical device if its intended use is to make a diagnosis or provide a treatment [20]. If the description or labeling of a mobile app specifically states that the primary purpose of the m-health app is not to make a diagnosis, provide treatment, or prevent a disease, the app does not qualify as a medical device [20] and hence does not need Health Canada approval.

1.5. Quality Control Measures on App Stores

Limited control measures exist over the quality of apps available on app stores. Google Play and iTunes have published guidelines on quality requirements for apps published on their app stores. Anyone with computer programming skills and a subscription to the developer programs of app markets can make an app and sell it on the app stores as long as the app meets the quality requirements set by the app stores [3]. According to Google and Apple design guidelines, the concept of *quality* refers to factors such as user interface, software functionality, and whether an app includes illegal content or content that infringes on intellectual property [21, 22]. The guidelines do not include factors such as the impact of apps on users' health or whether

the content of the app is based on research evidence. As a result, current quality requirements of the app stores do not meet the expectations of health care providers and users in terms of m-health apps.

1.6. Identifying High Quality Apps

In the absence of regulations and strict quality measures for m-health apps, consumers (i.e., patients and health care providers) may seek other sources for information to identify m-health apps with high quality [15]. Examples are the app descriptions on the app stores, ratings from other users, and endorsements made by users or health care professionals on the internet or social media [23].

1.6.1. App descriptions on the app stores

App developers typically include a short description of the content of their app on the download page of the app stores. Ideally, this description includes adequate information to help average users to decide if the app meets their needs. However, reviews of the app store content have shown that most of these descriptions are not trustworthy and are written in a way to convince users to download or purchase the apps. The app descriptions may include unrealistic or exaggerated claims about the effectiveness of the app or even incorrect information about the content of the app [2]. Therefore, researchers have recommended against judging the quality of m-health apps based on descriptions provided by app stores alone [11, 23, 24].

1.6.2. Users' ratings and reviews

App stores allow users to express their satisfaction with each app using a 5-star rating system; where five stars mean the user has been *very satisfied* with the app and no stars means that the user was *not satisfied at all*. Users have the option to post a comment about their experience using the app. Moreover, the app stores display the number of downloads for each

app, which some users may interpret as an indication of an app's popularity. App users tend to choose apps that are more popular based on number of downloads, higher star ratings, and positive reviews posted by previous users [25]. However, studies have shown that such indices may not be trustworthy. First, there have been reports among app developers' community about developers who pay for downloads or buy reviews and ratings [3]. App review services are so inexpensive that an app developer can pay $5 for 12 positive reviews for his product, or negative reviews for his competitor [2]. Second, studies have shown that even the endorsements made by real users provide little value to health literature and might be misleading. For example, in a review of apps for chronic conditions, Singh et al. [26] reported that consumers' ratings were poor indicators for app's clinical utility or usability. Studies have also shown that apps with lower ratings do not necessarily have poor quality [25]. Based on this evidence, the ratings and reviews provided by consumers are not recommended as a source to identify m-health apps with high quality.

1.6.3. Endorsements by health care professionals

The third source of information used by consumers to identify high-quality m-health apps is reviews published by health organizations or enthusiastic health care professionals. Some of these reviews come from formal sources such as government health organizations. An example of such formal app directories is the *National Health Services (NHS) Health Apps Library* of the United Kingdom. Currently, the library includes 11 categories of health apps: Cancer, Chronic Obstructive Pulmonary Disease, Dementia, Dental, Diabetes, Healthy Living, Learning Disabilities, Mental Health, Online Community, Pregnancy and Baby, and Other Conditions [27]. A team of NHS subject matter experts evaluates apps in the fields such as regulatory, clinical safety, and data security. Apps that meet the quality standards of NHS for clinical

effectiveness, safety, usability, and accessibility receive the NHS approved badge and are published on the NHS online app library.

Another example is the *Addiction and Mental Health Mobile Application Directory* by Alberta Health Services (AHS) in Canada [28]. The directory has grouped apps into eight major categories of disorders according to the Diagnostic and Statistical Manual of Mental Health Disorders (DSM-5). The directory is intended to serve as an information resource for researchers, health care providers, patients, and app developers. Also, the directory provides information on the app's developers, the app's operating system (i.e., iOS or Android), and the cost of each app. However, unlike the NHS library, apps on the AHS directory have not been evaluated against any quality criteria, and the directory provides no information on their usability or clinical effectiveness [28].

An example of nongovernmental app directories is the *practicalapps.ca* website. It is a joint project between the *Women's College Hospital Institute for Health System Solutions and Virtual Care* and the *Ontario Telemedicine Network* [29]. A group of volunteer family physicians with an interest in mobile health apps collaborate with this initiative. They review each app and rate its clinical validity, usability, privacy and security, accessibility, safety, and reliability [30]. Other examples of similar informal online reviews include *iMedicalApps*, which is a physician peer review mobile medical and health app website and *MedicalApp Journal* that uses Twitter and Facebook groups to peer review apps and provide information on the latest trends in digital medical technology [25].

Although online app directories provide a single point of reference for users to find m-health apps, health information experts have advised using such directories with caution. First, the trustworthiness of app directories, including those published by reputable sources, have been

questioned before. For example, the NHS health apps library was once shut down in 2015 after an independent study showed that 66% of the apps approved by NHS had serious security and privacy flaws [3]. In light of this report, NHS removed many of the apps that they previously considered as high quality before releasing the beta version of the library in April 2017 [27]. Second, ratings and reviews provided by these directories can quickly become obsolete due to the short life cycles of apps. M-health apps may become unavailable on the app stores within a year after their release. In a study of app store dynamics, Larsen et al. [31] demonstrated that every three days, one clinically relevant mental health app becomes unavailable to download from iTunes or Google Play Store. The most common reason is that the developers do not renew their accounts on the app stores, or they do not update the information on their apps. As a result, app directories need to frequently update their lists to keep up with the market, which costs considerable time and money.

1.7. M-health App Rating Scales

Recent literature supports a *bottom-up* strategy that emphasizes education of consumers about how to assess health apps [32] as a cost-effective alternative. One way to achieve this goal is to provide app consumers with rating scales that enable them independently to evaluate m-health apps based on core concepts and quality criteria for them to decide if an app meets their needs. Existing rating scales fall under two main categories: scales used to rate the *usability* of m-health apps, and scales to rate the *quality* of m-health apps.

1.7.1. App usability rating scales

Early scales developed to rate m-health apps focused on assessment of the usability of m-health apps. In the field of information technology, usability refers to the properties of an interface (e.g., a mobile app) that is associated with attributes such as ease of use, learnability,

efficiency, memorability, error rates and user satisfaction [33]. Most of the mobile usability rating scales are modified versions of earlier usability scales that were initially developed to rate the usability of computer programs and websites. An example of these scales is the *System Usability Scale (SUS)* developed by Brooke as "a quick and dirty usability scale" for information systems in general [34]. The scale has been used in many studies to evaluate the usability of m-health apps for different health issues such as weight loss [35], mobility [36], pneumonia [37], stress management [38], and adverse drug reaction [39].

Table 1.1 shows a list of scales, reported in the literature, used to rate the usability of m-health apps. The most common constructs used in these scales are perceived usefulness and perceived ease of use, adopted from the *Technology Acceptance Model (TAM)* [40].

Table 1. 1. App usability assessment scales

Name of the scale [References]	Scale construct(s)
Perceived Usefulness and Ease of use questionnaire (PUEU) [41]	Perceived usefulness, Perceived ease of use
Software Usability Measurement Inventory (SUMI) [42]	Efficiency, Affect, Helpfulness, Controllability, Learnability
System Usability Scale (SUS) [35, 38, 39, 43-52]	Ease of use, Functionality, Satisfaction
ISOMETRIC [53]	Suitability for the task, Self-descriptiveness, Controllability, Conformity with user expectations, Error tolerance, Suitability for individualization, Suitability for learning.
USE Questionnaire [54, 55]	Usefulness, Ease of use, Ease of learning, Satisfaction
Health-ITUES [56-58]	Impact, Perceived usefulness, Perceived ease of use, User control
Computer System Usability Questionnaire (CSUQ) [59]	Ease of use, User satisfaction, Errors, Functionality, Design

1.7.2. App quality rating scales

Although usability scales can rate the degree to which an m-health app is easy to learn and use, they do not cover criteria that are important to health care providers such as security of the app, customizability of the features, appropriateness to the users, and trustworthiness of the health information provided [60]. Therefore, attempts have been made to create quality rating scales for m-health apps to cover the criteria that are important to health care providers as well as usability experts. Some of these scales were developed to rate the quality of m-health apps that were specific to one health domain or health care issue. An example is *Enlight*, which was designed to rate the quality of apps used for behavior change. Another example is the *AQEL*, a scale for evaluating the quality of nutrition apps. (Table 1.2).

Table 1. 2. Quality rating scales for apps specific to a particular health domain

Name of the scale [References]	Scale construct(s)
APPLICATION scoring system [61, 62]	Comprehensiveness, Price, Paid subscription, Literature used, In-app purchase, Connectivity, Advertisement, Text search field, Inter-device compatibility, Other components (video, image, special features), Navigation ease, Subjective presentation
Quality Assessment tool for Evaluating Medical Apps (QAEM) [63]	Appropriateness of app, Reliability, Usability, Privacy
Enlight [64]	Usability, Visual Design, User Engagement, Content, Therapeutic Persuasiveness, Therapeutic Alliance, General Subjective evaluation of Program's Potential Credibility checklist, Evidence Based Program, Privacy explanation checklist, Basic security checklist
Nutrition App Quality Evaluation (AQEL) [65]	Behavior change potential, Support of knowledge acquisition, App function, Skill development, App purpose
MedAd-AppQ [66]	Content reliability, Feature usefulness, Feature convenience

Others were developed to rate the quality of health apps in general (Table 1.3). The most commonly used scale in this group is the *Mobile Application Rating Scale (MARS)*, used in more than 16 studies since 2015. *MARS* is a 23-item scale based on 372 criteria for evaluation and assessment of apps and web pages extracted from 25 published papers, conference proceedings, and online resources.

Table 1. 3. Quality rating scales for general m-health apps

Name of the scale [References]	Scale construct(s)
App Chronic Disease Checklist [67]	Engagement, Functionality, Ease of use, Information management
Health Care Apps Evaluation Tool [68]	Content (accuracy, understandability, objectivity), Interface design (consistency, the suitability of design, the accuracy of wording), Technology (security)
Royal College of Physicians' Health Informatics Unit checklist [69]	Who developed the app, and what is inside it? How well does the app work? Is there any evidence that the app does actually alleviate the problem?
The Organisation for the Review of Care and Health Applications- 24 Question Assessment (ORCHA-24) [70]	Data governance, Clinical efficacy, and assurance, User experience, and engagement
Mobile Application Rating Scale (MARS) [5, 71-86]	Engagement, Functionality, Aesthetics, Information, App subjective quality
Mobile Application Rating Scale-User version (U-MARS) [87]	Engagement, Functionality, Aesthetics, Information, App subjective quality

MARS has four objective quality subscales (engagement, functionality, aesthetics, and information), and one subjective quality subscale. Items in each subscale are rated based on a 5-point scale. *MARS* was first developed to evaluate the quality of mental health mobile apps on the iTunes app store [5]. The user version of *MARS* (*U-MARS*) was published one year later for

adolescents and young adults [87]. *U-MARS* is currently the only m-health app rating scale available for the general public to rate the quality of m-health apps.

1.7.3. Shortcomings of current app quality rating scales

While the current quality rating scales are important steps towards creating transparent and consistent methods to evaluate the quality of m-health apps, they have shortcomings. First, most of the current quality rating scales are not appropriate for the general public to use. Except for U-MARS, quality rating scales were initially designed for users with training in health or informatics, such as researchers and health care experts. The language used in these scales is technical and can be difficult for the general public. Moreover, the end users of m-health apps such as patients or older adults were not involved in the development process of any of these scales.

Second, none of these scales have been developed based on a theoretical framework for the evaluation of m-health apps. Instead, the developers have relied only on a "panel of experts" to determine the quality criteria measured by these scales. Without a theoretical framework to guide the assessment criteria used in quality rating scales, the validity of the ratings provided by such scales would be limited. In other words, app users cannot be confident if the rating scales include all assessment criteria that are relevant to the quality of m-health apps (content validity) or if the ratings provided by these scales truly measure the quality of m-health apps (construct validity).

The quality criteria measured by the current app quality rating scales are diverse. This diversity suggests a weak consensus among experts over the criteria to assess the quality of m-health apps. For example, *MARS* measures criteria such as engagement, functionality, aesthetics, and quality of information. On the other hand, *ORCHA-24* looks at data governance, clinical

efficacy and assurance, and user experience and engagement as the measures of quality. The *RCP* checklist only focuses on three factors: who developed the app, if the app works correctly, and if there is any evidence on the effectiveness of the intervention provided by the app. The diversity of the assessment criteria used in the existing app quality scales indicates that the concept of quality is a composite construct. However, none of the existing app quality scales capture the range of the assessment criteria to adequately measure the quality of m-health apps. Moreover, as stated, none of the current rating scales can support and justify the quality criteria they asses as valid measure of quality of m-health apps.

1.8. Frameworks for Evaluating M-Health Apps

Without a guiding framework, it is not possible to confidently define the assessment criteria that truly and adequately measure the quality of m-health apps. Earlier literature on frameworks for evaluation of technologies focused on predicting acceptance and use of information technologies in general. Because of the specific features of mobile health apps, in recent years, attempts have been made to create evaluation frameworks exclusively for mobile health applications. This section reviews the traditional technology acceptance and use models followed by the frameworks developed solely for the evaluation of mobile health applications.

1.8.1. Technology acceptance and use models

Successful implementation of technology solutions in health care depends on the acceptance of technology by patients, their family caregivers, and clinicians. Several theoretical models have been developed to explain the acceptance of technology (Table 1.4). Some focus on factors that impact the usability such as *Nielsen Model of Usability* [33], others focus on the acceptance of technology, such as *Technology Acceptance Model (TAM)* [88]. On the other hand, sociological research has taken a macro level approach that emphasizes the whole process from

purchase to acceptance and adoption of technology. An example of these models is Rogers' *Innovation Diffusion Theory (IDT)* [89] that explains the process through which users in a community accept an innovation. According to this theory, every successful innovation will be accepted first by enthusiasts (2.5% of population) followed by the visionaries (13.5%), the early majority (34%), the late majority (34%) and finally by the skeptics (16%) [89].

Table 1. 4. Technology acceptance and use models

Name of the model	Authors [References]	Criteria
Nielsen Model	Nielsen [33]	Usefulness, Utility, Usability, Easy to Learn, Efficient To Use, Easy to Remember, Few Errors, Subjectively Pleasing, Social Acceptability, Practical Acceptability, Cost, Compatibility, Reliability
TAM	Davis [40]	Social Influence, Perceived Ease of Use, Perceived Usefulness
UTAUT-2	Venkatesh, Thong, and Xu [90]	Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions, Hedonic Motivation, Price Value, Habit
STAM	Renaud and Biljon [91]	Social Influence, Perceived Usefulness, Facilitating Conditions, Ease of Use, Learning

Comprehensive frameworks also exist by merging earlier models (see Table 1.4). An example is the *Unified Theory of Acceptance and Use of Technology (UTAUT)* [92] that combines and unifies eight earlier models of technology acceptance and use. According to this framework, the mediator factors of technology acceptance are performance expectancy, effort

expectancy, social influence, and facilitating conditions. The effect of these factors can be moderated by other factors such as age, gender, and experience. A revised version of this model (UTAUT-2) also adds price value, hedonic motivation, and habit as mediators of technology acceptance and use [90].

Another example of comprehensive technology acceptance models is the *Senior Technology Acceptance and Adoption Model (STAM)* [91]. This model incorporates factors that influence technology acceptance and examines the process of technology adoption by older adults (i.e., merging TAM and IDT). According to this model, the adoption process has three phases: objectification, incorporation, and conversion or non-conversion. In the objectification phase, a technology user determines the role of technology in his or her life. During this process, the perceived usefulness of a technology combined with the social influence impact the user's behavioral intention to use the technology. In the second phase, through experimentation and exploration, the user evaluates the usefulness of the technology. Facilitating conditions and ease of learning and use can significantly impact the user's experience and eventually confirm or reject the usefulness of the technology. If the user confirms the usefulness of the technology, then, in phase three, the technology is accepted and adopted. If not, then the user rejects the technology.

Both *UTAUT-2* and *STAM* consider technology usefulness as an essential factor to predict the acceptance of the technology. Nielsen [93] defines usefulness as users' combined perception about usability and utility. He defines usability as "a quality attribute that assesses how easy user interfaces are to use" [93]. It has five components:

a) Learnability: how easy it is for users to accomplish basic tasks the first time they encounter the design?

b) Efficiency: Once users have learned the design, how quickly can users perform?
c) Memorability: When users return to the product after a period of not using it, how easily they can re-establish their proficiency?
d) Errors: how many errors do users make, how severe are these errors, and how easily can they recover from the errors?
e) Satisfaction: how pleasant is to use the design? [33]

Moreover, Nielsen defines utility as the extent to which the technology provides features that the user needs [33]. Nielsen, however, does not describe the elements of utility. According to Nielsen, usefulness is the extent to which technology meets a user's needs while at the same time, it is easy and pleasant to use.

1.8.2. Mobile app quality evaluation frameworks

Many studies have used rating scales based on technology acceptance models to assess m-health apps. For example, *System Usability Scale (SUS)* [34], which was developed based on *TAM*, has been used in at least 16 studies to evaluate apps on app stores [35, 38, 39, 43-52]. However, according to the recent studies, m-health app quality criteria that are important to consumers exceed those covered by earlier models of usability and technology acceptance. Examples of these criteria are the trustworthiness of the app developer, cost, and cultural appropriateness of the content [94], which were not covered in any of the usability models.

Several frameworks have been developed to meet these expectations regarding the evaluation of the quality of m-health apps. Examples of the most recent app evaluation frameworks are the *App synopsis* [95], *Health On the Net code* [96], *Canadian Framework for Assessment of Mental Health Apps* [94], and *Health Care Information* and *Management Systems Society (HIMSS)* framework [97] (Table 1. 5).

One of the earliest frameworks for evaluating the quality of apps was *App synopsis*, developed by Albrecht and Pramann [95]. They developed this framework based on conventional models of evaluating transparency and objectivity of clinical research. The primary purpose of this framework was to set reporting standards for app developers and encourage them to list features of m-health apps in a way that assist users in deciding which apps to choose objectively. Albrecht and Pramann suggest that app developers should be held responsible for describing their apps according to the standards set by app-synopsis through distribution channels such as app stores and corresponding web pages.

Table 1. 5. Frameworks for evaluating wellness apps

Name of App Evaluation Framework [References]	**Author(s)**	**Criteria**
App synopsis [95]	Albrecht, Noll, Von Jan, Jungnickel, and Pramann	1) Imprint (information about manufacturer) 2) Rationale (app's intended purpose); 3) Functionality 4) Validity and reliability of the information 5) Data requisitioning and management 6) Data protection 7) Data transmission
HON Foundation CODE [96]	Health on the Net Foundation	1) Information is Authoritative (the author is clear) 2) Purpose of the app is clear; Confidentiality 3) Information is referenced 4) Justification of claims 5) Developers contact details 6) Funding source is clear 7) Editorial content and advertisements are distinguished from one another

Table 1.5. Continued

Name of App Evaluation Framework [References]	Author(s)	Criteria
Canadian Framework for Mental Health Apps [94]	CIHR and MHCC	1) Effectiveness 2) Clinical claims 3) Usability 4) User Desirability 5) Security 6) Functionality 7) Interoperability 8) Supported Platforms 9) Target Users 10) Price 11) Transparency 12) Inclusion
Health Care Information and Management Systems Society (HIMSS) [97]	Chan, Torous, Hinton, and Yellowlees	A) Usefulness 1) Validity and accuracy (of app function) 2) Reliability (of app functions) 3) Clinical Effectiveness 4) Time and number of sessions B) Usability 1) Satisfaction 2) Usability (easy to use) 3) Disability accessibility 4) Cultural acceptability Socioeconomic accessibility C: Integration and Infrastructure: 1) Security Workflow 2) Data integration 3) Safety 4) Privacy

The App-Synopsis includes the following criteria [95]:

1) Imprint: the app should consist of information about the manufacturer and distributor and their associates

2) Rational: apps' intended purpose, target users, and its categorization as a medical app

3) Functionality: includes a description of the functionality and features of the app, details about the measures of the usability of the app

4) Validity and reliability: which information sources the app uses and how reliable these sources are

5) Data requisition and management: Description of the amount and types of data that are being collected and processed

6) Data protection and privacy: How the manufacturer adheres to data protection and privacy laws and regulations in the related jurisdiction

7) Data transmission and storage: description of all measures taken to protect data that is transmitted or stored by the app

The developers of app-synopsis suggest that the criteria can serve as a guideline for users, researchers, and app developers to test the quality of m-health apps [95].

Ben-Mussa and Paget [98] developed a framework by adopting quality criteria from Health on the Net Foundation code (HON-Code) to evaluate the most popular health mobile apps in the UK. HON foundation is an international not-for-profit accreditation program for health websites [99]. The foundation has set principles and standards for reliable and trustworthy health information on the World Wide Web. HON mandate is to reduce misunderstanding by clarifying the context of the information and enable users to validate and choose suitable information for

themselves [98]. Ben-Mussa and Paget framework includes eight principles from the HON-Code:

1) The information must be authoritative: All medical information presented by (or calculations performed by an app) must be attributed to an author, and his or her training in the field must be mentioned.

2) Purpose (of the app): A statement declaring that the (app) is not meant to replace the advice of a health professional is necessary. A brief description of the (app)'s mission, purpose, and the intended audience is required. A summary of the organization behind the (app), its mission, and its use is also required.

3) Confidentiality: The (app publisher) must describe its privacy policy regarding how you treat confidential, private, or semi-private information such as email addresses and the content of emails received from or sent to (its users).

4) The information must be documented, referenced, and dated: All medical content (including calculations and formulae) must have a specific date of creation and the last modification date.

5) Justification of claims: All claims, including information about the benefits or performance of any treatment (medical or surgical), commercial product, or service, have to be backed up with scientific evidence such as medical journals, reports, or others.

6) App contact details: The (app) must be operational, and the information must be accessible and presented clearly. There must be a way to contact the app publisher, such as a working email address or contact form, for visitors who would like to have more details or support.

7) Funding: The app publisher must include a statement declaring its sources of funding.

8) Editorial and advertising policy: App developers must declare any conflicts of interest and external influences which could affect the objectivity of the editorial content in the disclaimer. All apps that display paying banners must have an advertising policy. This policy must explain how the publisher distinguishes between editorial and advertising content and which advertisements are accepted [98].

Chan et al. [97] have developed an evaluation framework for mental health mobile apps following the quality evaluation guidelines from the Health Care Information and Management Systems Society (HIMSS), the American telemedicine association, and the U.S. Federal government. This framework proposes three dimensions of evaluation criteria for mental health mobile apps, which are usefulness dimension, usability dimension, and integration and infrastructure dimension.

The usefulness dimension includes the following factors:

1) Validity and accuracy: if the app works as advertised;
2) Reliability: if the app functions consistently;
3) Effectiveness: if the app is clinically effective;
4) Time and number of sessions: how much time should a user spend on the app to see benefits).

The usability dimension includes the following:

1) Satisfaction and rewards: if the app is pleasurable
2) Usability: can the user easily use the app
3) Disability accessibility: if the app incorporates the needs of blind users
4) Cultural accessibility: if factors such as language and ethnicity define the app

5) Socioeconomic accessibility: if the app takes into consideration age, income, and tech literacy)

Finally, the Integration and infrastructure dimension includes the following:

1) Security: if the data is encrypted
2) Workflow integration: if the app works within the user's workflow
3) Data integration: if the app shares data with other apps or systems
4) Safety: if the app takes into consideration patient safety for example risk of suicide
5) Privacy: if the app contains a robust privacy policy that addresses the type of information collected, stored, and shared [97]

The most recent framework for the evaluation of mobile apps has been developed by the Mental Health Commission of Canada (MHCC) and the Canadian Institute of Health Research (CIHR). The main purpose of this framework was to help m-health app users across Canada make informed decisions about mental health apps that they use. Users may include people who manage their mental health, health care providers who are interested in recommending good apps to their patients, and app developers who like to improve their products [100]. Stakeholders such as app users and developers, health care providers, mental health advocates, people with lived experience of mental health problems, policymakers, and researchers participated in several focus groups to contribute to the development of this framework [100]. The framework consists of two major criteria: (1) evaluation, and (2) information. The evaluation criteria consist of the following factors [100]:

1) Effectiveness: what is the app's intended purpose? Can it do what it says it will? Is there proof?

2) Clinical claims: If the app makes certain clinical claims (e.g., reduce stress or anxiety) does it give proof of its efficacy?
3) Usability: Is the app user-friendly and engaging enough to make people want to keep using it?
4) User desirability: Will the target users want to (or are able) to use the app?
5) Security and privacy: Does the app clearly state how it will collect, store, use, and protect personal health information? Is this information easy to find or hidden deep within the app? Does the app meet all applicable federal and provincial/territorial legislative standards and requirements regarding personal health information?

The information section of the framework consists of the following criteria [100]:

1) Functionality: What functions does the app offer? (e.g., journaling, mood tracking, guided exercises)
2) Interoperability: Does the app use open standards allowing it to exchange data with other health apps or tools?
3) Support platform: Is the app exclusive to one platform which may create accessibility barriers? Or is it available to many users across Android and iOS and other devices?
4) Target users: Who is the intended audience for the app? Is it clear who should or should not be using it?
5) Price: Is the app upfront about its costs, or are there hidden or extra fees? Will the price create accessibility barriers for the intended users?
6) Transparency: Does the app clearly state the individuals or organizations involved in its development? Does it state who provided the funding for its development?

7) Inclusion: Where the target users involved in the development and testing of the app to ensure it responds to their needs and expectations? How diverse was the user input? Were people from a variety of populations with unique mental health challenges involved (e.g., immigrant, refugee, ethno-cultural and racialized communities, First Nations, Inuit, Metis, LGBTQ2+, homeless people, seniors, youth)?

1.9. Summary and Gaps

The number of m-health apps is rapidly increasing in app stores. In addition, the general public's interest to use m-health apps as an adjunct to conventional health practices is increasing. However, because of the inadequate quality control mechanisms on app stores as well as lack of stringent regulations for m-health apps, there is a risk of harmful apps or apps with inferior quality available to the public. In the absence of recommendations by health experts, users may choose apps based on unreliable information such as reviews and ratings on the app download page or the number of downloads for an app.

Government and private organizations have made efforts to review m-health apps and introduce the ones with high quality to the public through online directories of apps. However, due to the short life cycle of m-health apps, the maintenance of such directories is costly and impractical. The ratings on these directories may quickly become obsolete because by the time the review process is complete a new version of the app may appear on the store or the publisher may remove the app from the app store. A more realistic alternative is to create a criteria-based scale for users and their advocates to evaluate the quality of apps based on their needs.

Several m-health app rating scales exist for experts and researchers. Most of them are long and complicated scales that are not usable by the general public who have little or no formal

training in health or informatics. None have been developed using a theoretical framework although several frameworks exist. Without using a theoretical framework it is not clear if the existing app quality rating scales capture the criteria to adequately assess the quality of m-health apps. There is a need for a rating scale based on theoretical frameworks of m-health app evaluation that can be used by expert and non-expert users and can help them to make an informed decision when selecting and using an m-health app.

1.10. Rationale

The review of the current state of knowledge shows that in the absence of formal quality control on mobile health applications (m-health apps), end users may access apps that do not meet their needs. Existing scales are relatively long and complicated for average end users who have limited or no training. Moreover, none of these scales have been based on theoretical frameworks of app evaluation to support criteria that define quality. There is a need for a short, easy-to-use, and user-friendly rating scale or index for the general public users of m-health apps (including older adults) that is theoretically supported by frameworks of app evaluation and models of technology acceptance and use.

1.11. Objective

The objective of this study was to create an assessment tool based on frameworks of app evaluation and users' experience that helps m-health app consumers (e.g., patients and family and professional caregivers) rate the quality of m-health apps and make an informed decision when selecting m-health apps. This assessment tool is named the Alberta Rating Index for Apps (ARIA).

1.12. Research Questions

This study addressed the following research questions:

1) What are the criteria to rate the quality of mobile health apps?
2) How can users of m-health apps measure the extent to which m-health apps meet the quality criteria mentioned in question one?
3) Can health care providers, older adults, and adult with a chronic health condition use ARIA as a criteria-based index to reliably rate the quality of m-health apps?
4) What is the criterion-based validity of total quality scores of ARIA compared to U-MARS?

CHAPTER 2: METHODS

A multiple strategy sequential design with qualitative and quantitative components was used to achieve the objectives of this study and answer the research questions. Following Streiner and Norman's guidelines for developing health measurement scales [101], the investigator conducted this study in three phases:

Phase 1: To identify relevant assessment criteria for quality of m-health apps, and generate an item pool (research question 1)

Phase 2: To validate the content of the item pool and shortlist the items (research question 2).

Phase 3: To develop and test reliability and criterion related validity of ARIA (research questions 3 and 4)

2.1. Phase One – Identifying the Quality Criteria and Generating an Item Pool

Phase 1 of the study consisted of two main research activities. In the first research activity, the investigator used focus groups of users of m-health apps to identify the criteria to rate the quality of m-heath apps from users' perspective. In the second research activity the investigator developed an item pool to measure the quality criteria identified in the first research activity.

2.1.1. Identifying the relevant quality criteria for m-health apps according to end users

To identify the criteria relevant to the quality of m-health apps, the investigator first extracted the quality criteria from models of technology acceptance and use and frameworks for app evaluation. Next, the investigator examined which of these criteria were important from end users' perspective. To achieve this goal, he conducted six focus groups with m-health app consumers and app developers. Focus groups are interviews with small groups of people on a

specific topic that help researchers to get a variety of perspectives and increase confidence in patterns that emerge from the interviews. Focus groups are widely used in marketing and business research that aim to evaluate the opinions of consumers about a specific product or service [102].

2.1.1.1. Sampling and inclusion criteria

The investigator used a convenience sampling method to recruit the following groups of participants for this study: older adults, adults with a mental health condition, health care providers, and app developers. This strategy is useful in focus groups to "create specific information-rich groups that can reveal and illuminate important group patterns" [102].

The investigator recruited participants using the following inclusion criteria:

1) Able to communicate in English (verbal and written)
2) Owns and uses a smartphone

In addition, participants also met the following group-specific inclusion criteria:

1) Adults with a mental health condition
    a) Aged 18 to 64 years
    b) Have a history of a mental health condition

This group provided insight on if and how adults aging with a mental health condition use mobile apps to manage their wellness.

2) Healthy older adults
    a) At least 65 years of age

This group provided insight into the age-related needs of older adults who are interested in using m-health apps.

3) Health care providers
    a) Be a licensed health care provider
    b) Have at least two years of experience providing health services

The investigator included this group to investigate if health care providers use m-health apps in their practice, if their patients ask them to recommend any m-health apps, and how they choose apps to use or recommend to their patients.

4) App developers
    a) Be experienced with the development of applications related to health care
    b) Hold a degree certificate or a diploma from a university or college program in the field of computer science or computer engineering

This group provided insight on what factors contribute to the interface, technical functionality, and security and privacy of m-health apps.

2.1.1.2. Exclusion criteria

Participants were excluded from the study if they:

1) were not familiar with smartphones
2) had severe physical, visual, and hearing limitations that could not be corrected with the use of an assistive device
3) had severe mental or cognitive impairments who were not able to provide informed consent

2.1.1.3. Procedure

During each focus group, the investigator (PA) assumed the role of the facilitator. He explained the study objectives and procedure to the participants. Participants read the

information letter (Appendix B), asked any questions that they had about the study, and signed the consent forms to participate in the focus group.

During the focus group, the investigator first gave each participant a tablet on which he had installed three mental health apps earlier. Participants were requested to use each app for at least 5 minutes to become familiar with the type of apps that were of interest for the study. In the next step, the investigator asked participants to suggest items that they believed should be on a scale or index to rate the quality of mental health apps. Next, the facilitator requested the participants to discuss the quality criteria that they believed were important to assess the quality of mental health apps. The facilitator took notes of the criteria mentioned by the participants at each focus group.

The facilitator also provided the list of the criteria he identified earlier from the technology acceptance models and app evaluation frameworks to the participants and asked them to discuss the importance of these criteria. The criteria were appearance, compatibility, security, cost, ease of use, enjoyability, usefulness, learnability, privacy, social acceptability, and recovery from errors. Next the facilitator asked participants to rank the criteria from the technology acceptance models and app evaluation frameworks from the most important to the least important. The facilitator audio recorded the focus groups digitally. Each focus group lasted for a maximum of two hours, with a 15-minute break in the middle of each session.

The initial plan was to invite participants to a second focus group to classify the items that they had suggested. However, none of the participants in any of the focus groups suggested items to rate the quality criteria for mobile apps. For this reason, the investigator did not ask them to attend a second focus group. Instead, the investigator used previously developed app quality scales as the main source to generate the item pool.

2.1.1.4. Data analysis

The investigator transcribed all the recordings verbatim. Next, a researcher colleague of the investigator (Noelannah Neubauer) randomly selected 50% of the transcripts and verified their correctness by listening to the original audio recordings. The investigator used thematic analysis method as described by Braun and Clarke [103] to identify common themes on quality criteria mentioned by different groups of participants. Thematic analysis is "a method used for identifying, analyzing, and reporting patterns (themes) within data" [103]. Unlike other methods of qualitative analysis such as grounded theory, thematic analysis is not bounded by any pre-existing theoretical position, and for this reason, it has more flexibility in terms of ontology and epistemology that makes it suitable for different methodologies [103].

To analyze the data, the investigator entered the transcripts into QSR International's NVivo 12 (2018) Data management Software. The investigator read the transcripts several times to become familiar with the data. In the next step, he systematically coded the features of data that were of interest according to the research question (i.e., factors to evaluate the quality of m-health apps, the interpretation of participants of the identified criteria, and the importance of each quality criteria). Next, relevant codes were collated to generate themes (criteria to assess the quality of m-health apps). A second independent researcher colleague (Christine Daum) reviewed the codes in one-third of transcripts that were randomly selected to verify the validity of the coding process.

2.1.2. Generating the item pool

In the second research activity of phase one, after identifying the quality criteria for m-health apps, the investigator generated an item pool using the previously developed scales for

mobile apps as a primary source of items. The investigator identified these scales through a systematic review of the literature on scales for mobile apps.

2.1.2.1. Data source and search strategy

The investigator examined the peer-reviewed literature between Jan 2000 and July 2018 in three medical databases (Medline and CINAHL via EBSCO and PsycINFO via Ovid) and one computer science database (IEEE Explore). The medical databases were searched using the following search string that was developed using MeSH terms (where applicable) and keywords in consultation with a subject librarian with expertise in the area of rehabilitation medicine:

(“Mobile Application*” or “mobile app*” or "Portable Electronic App*" or "Portable Software App*" or portable software app*) AND (quality or criteria or evaluat* or usability).

The Computer science database was searched using the following terms: ((Mobile application*) AND (quality OR usability)) AND (Health).

The grey literature was searched through the following online sources using “health mobile app” as our search string: Nielsen Norman Group website, Google material design guidelines, Health on the NET foundation (HON) website, and Google custom search for Canadian government documents.

2.1.2.2. Inclusion criteria

Studies that met the following criteria were included in the review.

1) Were published from 01/01/2000 and 31/07/2018
2) Were published in any language and were available in full text in peer-reviewed journals or conference proceedings from searched databases and the grey literature
3) Used any design (descriptive, experimental, reviews) or any methodology (qualitative, quantitative, or mixed)

4) Reported a scale or questionnaire for rating the usability or quality of health-related mobile or tablet applications

2.1.2.3. Exclusion criteria

Studies that met the following criteria were not included in the review.

1) Either the abstract or full text was not available
2) Were about apps not related to health or wellness (e.g., business apps, educational apps, etc.)
3) Did not report any scale for rating the usability or quality of health-related mobile or tablet applications
4) Were only about websites and not portable device apps

2.1.2.4. Study selection procedure

The investigator exported all retrieved results to a reference manager software (EndNote X7.7.1, © 1988-2016 Thompson Reuters) and removed all duplicates. Next, the investigator and one of his research colleagues (Noelannah Neubauer) first independently screened the titles and abstracts to select potential papers for full-text review based on the selection criteria. In the next step, both reviewers read the full texts of selected articles and identified the ones that contained a scale for evaluation of m-health apps and chose them for data extraction. At the end of each step, the reviewers resolved disagreements through consensus. Unresolved disputes were forwarded to a third colleague (Antonio Miguel Cruz) to settle.

2.1.2.5. Classification of scale items

The investigator followed the methodology used by Nouri et al. [104] for this step. The investigator extracted the items identified from the scales and entered them in an excel spreadsheet. Next, the investigator compared similar items from different scales and grouped

them under relevant quality criteria identified from the focus groups in the previous phase or the frameworks of app evaluation discussed in chapter one. Examples of quality criteria used were usefulness, ease of use, enjoyability, cost, security, privacy, and aesthetics.

In the next step, the investigator forwarded the categorized items to a panel of experts, including three researchers with experience in evaluation of health technology and scale development (Adriana Rios Rincon, Antonio Miguel Cruz, and Mary Roduta Roberts). The panel members reviewed the items independently and rated the degree to which they agreed with the classification made by the investigator, using a 3-point Likert scale (1 – the item does not align at all with suggested quality criterion; 2 – the item aligns moderately with the suggested quality criterion, and 3 – the item aligns perfectly with the suggested quality criterion).

For each item, the investigator accepted the majority of votes between the panel members (2 out of 3). When there was not a majority vote, the investigator accepted decision by Antonio Miguel-Cruz since he had more experience with technology acceptance models and usability studies. In the final step, the investigator excluded the redundant items and items that the panel did not identify as relevant. The investigator edited the remaining items to improve readability and used them to generate the item pool.

2.2. Phase Two – Developing ARIA

In phase two, first the investigator validated the content of the item pool using an online survey method and calculating content validity index for each item. Next, the investigator recruited a panel of experts to short list the items that were identified as valid in the online survey and develop the pilot draft of the index, which had an acceptable length (number of items) for the general users, and covered all quality criteria for mobile health apps.

2.2.1. Content Validation

In the first research activity, users and developers of m-health apps reviewed the relevance and adequacy of items per each quality criteria. This process is technically referred to as *content validation* [101].

2.2.1.1. Participants

Participants were selected based on the same inclusion and exclusion criteria used to recruit focus group participants in phase one. Following Lynn [105] recommendations for content validation of the scale items, a minimum of 10 participants would be adequate to rate the content validity of a scale. Participants were recruited using a convenient sampling method and through the personal contacts of the investigator, participants in the focus groups in phase one, and from a network of mental health research collaborators affiliated to the Mental Health Commission of Canada.

2.2.1.2. Procedure

An online survey was developed using the Survey Gizmo Platform based on generated items and relevant quality criteria. The link of the online survey was forwarded to each participant by e-mail. A paper version of the survey is provided in Appendix C. Next, participants rated the relevance of each item to the given quality criterion on a 3-point scale: 1- Not relevant, 2- Relevant and 3- Unable to assess relevance. Moreover, participants were requested to add any additional items that they believed were missing from the list and suggest any revisions to the wording of the items.

2.2.1.3. Data analysis

The investigator calculated the content validity index (CVI) for each item by dividing the number of participants who endorsed the item as relevant to the total number of participants in

the survey. To avoid inflation of agreement due to chance factors Lynn's criteria was used for acceptability of CVI that incorporates the standard error of the proportion [105]. According to Lynn, the acceptable CVI for each item when 10 judges participate in ratings is 0.8 [105] (i.e., at least 80% of the participants must endorse the item as relevant).

2.2.2. Shortlisting the validated items

The objective of the second research activity in phase two was to reduce the length of the index by removing redundant or less relevant items to make the index acceptable to the general public while maintaining the content coverage of the index (i.e., cover all relevant criteria for quality of m-health apps). To achieve this goal, the investigator conducted a focus group of experts and stakeholders to review the items and shortlist the most relevant ones through discussion and voting.

2.2.2.1. Participants

The investigator recruited the participants based on the following inclusion criteria:

1) At least 18 years of age
2) Able to communicate in English (verbal and written)
3) Own and use a smartphone for at least one year

Participants were purposefully selected from the following stakeholder groups so that at least one older adult (>60 years of age), one adult with lived experience of mental health conditions, one mental health care provider, and one person with expertise in technology usability evaluation would be present in the focus group.

2.2.2.2. Procedure

Participants received the study information letter and consent form (Appendix B) either in person or by e-mail. The investigator answered any questions they had and collected their

signed consent to participate in the study (in person or via e-mail). After providing consent to partake in the study, participants received the study package that included the agenda for the focus group and a copy of the index review form (Appendix D), at least one week before the focus group.

The index review form included 51 items distributed unevenly under 10 quality criteria for mobile health apps. Participants were requested to review the items and indicate which items should be kept and which ones should be removed from the index by marking "X" in a box provided on the review form in front of each item. Participants were reminded that the objective was to reduce the length of the index while it covered all 10 quality criteria. Participants were also requested to use the comment section to explain their decisions or make any suggestions.

On the day of the focus group, the investigator assumed the role of the group facilitator. First, each participant named items she or he had chosen to keep and which ones to remove with a brief justification for her or his choice. Next, participants voted on which items should be kept for each criterion. Any disagreements were solved through discussion until consensus was achieved. At the end of this phase, the investigator used the shortlisted items to develop the pilot draft of Alberta Rating Index for APPS (ARIA) (Appendix E).

2.3. Phase Three – Testing ARIA

Phase three consisted of two research activities. In the first activity, the pilot draft of the index was tested in a pilot study for readability and comprehensibility of the items. In the second activity, the reliability and validity of the index were examined.

2.3.1. Pilot Testing

To investigate the readability and comprehensibility of the items, the investigator tested the first draft of ARIA in a pilot study using a "think aloud" method [101]. In this method,

participants used ARIA to rate the quality of an m-health app and think aloud as they answered the questions in the index.

2.3.1.1. Participants

Studies recommend a minimum of five participants for pilot studies that use the think-aloud method [33]. Participants were recruited purposefully based on the inclusion and exclusion criteria used in 2.2.2.1.

2.3.1.2. Procedure

The investigator met with each participant in person at a location that was convenient to the participant. The investigator presented each participant an information letter (Appendix B) and obtained participants' signed consent to partake in the study. Next, each participant downloaded a sample mental health app (i.e., CBT-i Coach) from iTunes or Google Play store on her or his smart phone. This app was chosen purposefully from the mental health app directory of Alberta Health Services (AHS) [28] because of its features that applied to our index. Next, each participant used the pilot draft of ARIA (Appendix E) to rate the quality of the CBT-I Coach app. Participants were requested to think aloud and verbalize their thought process while they responded to each item in the index. Participants were asked to explain the following:

1) how they interpreted each item
2) if the items were easy to understand
3) if the scale responses were appropriate for the items
4) if they needed more prompts to answers each item
5) if they recommended any revisions to improve the comprehensibility of the items

The investigator recorded each session digitally and wrote any comments or recommendations made by the participants. The investigator revised ARIA based on the comments and suggestions made by the participants and used it in the next phase of the study.

2.3.2. Testing the reliability and validity of ARIA

In the second research activity of phase 3, the inter-rater reliability and criterion related validity of ARIA were studied. To test the inter-rater reliability, the investigator used a fully-crossed design within the framework of generalizability theory (G-Theory) [106]. In this design, raters were recruited from different groups of users of m-health apps. All raters rated all apps against all quality domains (raters x apps x quality domains) [106]. The fully crossed design enabled us to investigate the different sources of variability (i.e., raters, apps, and index items) that could contribute to the variability of app quality scores obtained from using the index (i.e., ARIA).

The investigator asked the raters to evaluate the quality of apps one more time with the user version of the Mobile Application Rating Scale (U-MARS) [87]. The total scores from U-MARS were correlated later with total scores from the new index to test the criterion-related validity of the index. Although U-MARS is not the gold-standard scale for rating mobile apps, it is currently the only scale designed for end users to rate apps.

2.3.2.1. Participants

Participants (hereafter referred to as raters) were recruited from potential end users of m-health apps based on the following inclusion criteria:

1) At least 18 years of age
2) Able to communicate in English (verbal and written)
3) Own and use a smartphone for at least one year

4) Half of the raters were iPhone users, and the other half were Android device users.

At least two raters were needed to estimate the rater variance component using G-theory. Moreover, the reliability of scores across three groups of raters (older adults, younger adults, and health care providers) was of interest to the investigator. For this reason, 12 raters were recruited for this study (4 older adults, 4 adults with lived experience of mental health condition, and 4 mental health care providers). Two raters in each group were iPhone users, and the other two were Android users.

Participants were excluded if they:

1) Had severe physical, visual, and hearing limitations that could not be accommodated by the use of an assistive device.
2) Had severe mental or cognitive impairments who were not able to provide informed consent.

2.3.2.2. M-health apps

The purpose of the new index was to help care providers and adults with mental health conditions identify mental health apps with high quality and usability. Mental health apps are under different categories on app stores, including wellness, medical, and lifestyle. However, due to limitations of the search function of app stores, searching for mental health apps on the app stores may not identify a representative sample of apps that span various mental health conditions and target audience. As an alternative, the investigator used the Alberta Health Service's (AHS) *Addiction and Mental Health Mobile App Directory* [28] as the source of m-health apps in this study.

The AHS m-health app directory includes a variety of apps for different populations and mental health conditions. The directory has been developed through a comprehensive search for

apps in different databases, organization websites, news articles, and information sources in the public domain. For this reason, the investigator identified the directory as a better source to select a representative sample of mental health apps that were likely to be used by patients or recommended by health care providers. The directory had 39 apps under several categories of mental health conditions [28]. However, some of the apps on the directory were not suitable for the target users of our index. Examples were apps for neurodevelopmental conditions and apps developed as an educational tool for medical students.

According to the developers of the directory, the quality and usability of the apps on the directory were not evaluated. Therefore, to include apps that were relevant and to avoid selecting apps that were potentially harmful to the participants, the investigator pre-screened and shortlisted the apps on the directory based on the following inclusion criteria:

1) Available on both iOS and Android markets
2) Developed for general app users with no medical training
3) Designed for mental health conditions common in adults and older adults
4) Did not include inappropriate or potentially harmful content

Fifteen apps were shortlisted in the following categories:

1) Substance abuse (3 apps),
2) Mood disorders (2 apps),
3) Anxiety disorders (2 apps),
4) PTSD (1 app),
5) Eating disorder (1 app),
6) Stress (3 apps),
7) Medication reminder (2 apps)

8) Sleep (1 app)

Next, the investigator selected at least one app from each category to represent all relevant mental health conditions covered in the directory. For categories with more than one app, the selection was made by flipping a coin. The investigator selected two apps from the stress management and Substance abuse categories since there were three apps under each. This process resulted in a sample of 11 apps that were relevant to mental health conditions and were accessible on both iOS and Android app markets.

2.3.2.3. Procedure

The investigator first met with each rater individually and presented the information letter (Appendix B) and obtained their signed consent. Next, the investigator demonstrated for each rater how to use ARIA and the user version of the mobile application rating scale (U-MARS) [87] to rate the quality of mobile apps. Each rater used ARIA and U-MARS in a practice round using two apps (selected from the directory in addition to the sample of this study). The investigator was present during the practice rounds. Raters had the opportunity to ask any questions until they felt comfortable to use both ARIA and U-MARS.

Next, each rater received the list of 11 apps (numbered as app number 1 to 11). Also, each rater received adequate paper copies of ARIA and U-MARS (Appendices F-H). Raters took the list of the apps and the scales home to rate the apps on their own. Raters were instructed to download each app on their smartphone and use it for as long as they needed to become familiar with the app's features. This time could vary for different apps. Next, raters were instructed to rate apps numbered 1 to 5 with ARIA first and U-MARS second; and apps numbered 6 to 11 with U-MARS first and with ARIA second, to follow a crossover design and prevent the impact of fatigue, learning effect, or carryover effect.

The investigator instructed the participants in the health care provider group to rate the apps based on a detailed mock patient profile (Appendix I). Participants in the older adults and adults with a mental health condition groups were instructed to rate the apps according to their needs and expectations (i.e., as if they wanted to use the apps themselves). For this reason, the results were mainly analyzed by groups collapsed across raters.

The raters had 30 days to complete the ratings for all 11 apps. During this time, they could contact the investigator if they had any issues or questions. Besides, the investigator gave each rater a follow-up call every week and asked if they had any questions or concerns. After 30 days, the investigator met with each rater and collected completed copies of the scales.

2.3.2.4. Data analysis

Descriptive analyses (frequencies, means, and standard deviation) were used to describe the distribution of app total scores by raters within each group. Moreover we explored if a floor effect (percent of apps rated at the floor of the index, equivalent of 10% worst possible score on the index) and a ceiling effect (percent of apps rated at the ceiling of the scale, equivalent of 10% best results on the index) were present [107]. Next, two-way 3 x 4 (Group x Rater) mixed factorial analysis of variance (ANOVA) was conducted to investigate:

1) The between subject effect: If the average total index score given by each group of raters was significantly different from the other groups (between subject factor with three levels: health care providers, older adults, and adults with mental health condition).
2) The simple effect of within subject effect: If raters within each group, rated the apps differently from each other (within subject factor with four levels of raters).

In the next step, item level analysis was conducted. Descriptive analyses (frequencies and mean) were used to investigate the distribution of responses to each item by raters within each

group. The inter-item correlation and item total correlation analysis was conducted using the Spearman rank's correlation test. The expected correlation between items that measured the same criterion was moderate to high. On the other hand, items that measured different quality criteria were not expected to correlate highly with each other. All items were expected to correlate moderately with the index total score [101].

A two-facet generalizability study (G-study) [106] was conducted with quality of apps fully crossed with index items and raters (i.e., A x I x R), for each rater group separately (i.e., health care providers, older adults, and adults with mental health condition). This study investigated the proportion of variance associated with index items, raters, and quality of apps, and calculated the G-coefficient as the measure of reliability for the item scores.

Next, the investigator conducted a single-facet generalizability (G-study) [106] with quality of apps fully crossed with raters (i.e., App x R) for each rater group separately. This study investigated the proportion of variance associated with quality of apps and raters, as well as the G-coefficient as the measure of reliability for the total index score.

Then, to test the criterion-related validity of ARIA, the investigator correlated the total scores from ARIA with the total scores from U-MARS for the same apps.

IBM SPSS statistical package 25 [108] was used for all the statistical analyses. A peer-reviewed SPSS Syntax [109] was used to estimate the variance components and the G-coefficient in the G-Study. The type one error (alpha) was set at 0.05 for all the analyses.

Finally, raters provided a general feedback in an exit interview regarding their experience of using ARIA in comparison to U-MARS (i.e., which one was easier or more useful). The interviewer recorded raters' feedback digitally and transcribed the recordings verbatim. The

investigator summarized the feedback and presented them in the final section of chapter three of this dissertation.

2.4. Ethical Considerations

The methods and procedures of all the phases in this study were approved by the Ethics Review Board of the University of Alberta (REB2) on February 9, 2017, protocol number: Pro00067856 (Appendix A).

CHAPTER 3: RESULTS

The objective of this study was to create an assessment tool to assist end users to make informed decisions when selecting an m-health app. These end users include older adults, patients, family caregivers, and health care providers. To achieve this objective, a multi strategy study was conducted in three phases. Each phase consists of a number of research activities. In this chapter, the findings of the research activities within each phase and how they connect to each other are presented.

3.1. Phase One – Identifying the Quality Criteria and Generating an Item Pool

In phase one, the investigator conducted two research activities. First, the quality criteria that were important to m-health app users (i.e., older adults, patients, and health care providers) and app developers were identified. In the second activity, an item pool was generated through a systematic review of the m-health app quality rating scales developed in previous studies.

3.1.1. Quality criteria for m-health apps according to end users

In the first research activity in phase one, the criteria that were important to users of mobile health apps were identified. First, the investigator extracted 11 quality criteria for mobile apps from technology acceptance models (i.e., UTAUT2, STAM, and Nielsen's Model) and frameworks for app evaluation (i.e., App synopsis, HIMSS, HON, and CIHR- MHCC). The criteria identified were appearance, compatibility, security, Affordability, ease of use, enjoyability, usefulness, learnability, privacy, social acceptability, and recovery from errors. In the next step, end users of m-health apps participated in six focus groups and discussed which of these criteria were important from their perspective. During the focus groups, participants discussed their experience with using m-health apps, types of m-health apps they used, and factors they considered important when selecting and using an app. The investigator also

presented the list of 11 quality criteria that he had identified earlier to the focus group participants and asked them to rank the criteria from the most important (1st place) to the least important (11th place).

3.1.1.1 Participants' characteristics

Twenty five end users of m-health apps (seven older adults, seven adults with mental health condition, and 11 health care providers) and seven app developers participated in six focus groups (Table 3.1) and discussed which of these criteria were important from their perspective.

Table 3. 1. Focus Group Participants

Groups	Gender (n) Male/Female	Mean Age (SD)	Length of time using a smartphone (years)
Older adults	2/5	73 (5.3)	At least 2
Adults with mental health condition	3/4	53 (8.9)	More than 5
Health care providers	4/7	38 (15)	More than 5
App Developers	4/3	29 (8.4)	More than 5

Seven older adults, five females and two males, participated in two focus groups. The first focus group was conducted at Seniors Association of Greater Edmonton (SAGE) and the second focus group was conducted at the Faculty of Rehabilitation Medicine, University of Alberta. All participants owned and used a smartphone for at least two years, however, two participants said that this technology was very new to them, and they were still learning how to use it. Four participants did not know how to download and install apps from app stores, and they relied on assistance from spouse or children to download new apps.

Seven adults with mental health condition, three females and four males, participated in one focus group. Participants were all members of the *Organization for Bipolar and Affective Disorder* (OBAD) in Calgary, where the focus group was conducted. All participants in this focus group owned and used a smartphone for more than five years.

The investigator recruited 11 health care providers, seven occupational therapists and four psychologists, who participated in two focus groups. The first focus group was at the Faculty of Rehabilitation Medicine (Corbett Hall), University of Alberta. Participants in the first focus group were occupational therapist researchers. The second focus group was at the Canadian Back Institute (CBI) Health center in Calgary. Participants in the second focus group were clinicians (four clinical psychologists and three occupational therapists).

The last focus group was held at the Department of Computing Science (Athabasca Hall), University of Alberta. Participants in this focus group were graduate students and post-doctoral fellows (three females and four males) who had experience in software and app development. This group was referred to as the app developer group in this dissertation.

3.1.1.2 Important quality criteria from the app user's perspective

Participants in each focus group discussed the quality criteria that were important to them. However, they were not able to suggest many criteria on their own. In order to facilitate the discussion, the investigator presented them the framework he developed from the models of usability and frameworks for evaluation of app quality. Next, the investigator asked the participants to discuss the importance of the criteria presented to them and rank them from the most important to the least important.

Table 3.2 shows the rankings made by different participant groups. The most important ones across all groups of participants were usefulness, ease of use, and privacy and security. Privacy and security were ranked almost consistently together, hence they are presented in the same cell in Table 3.2. On the other hand, participants ranked social acceptability and compatibility at the bottom of the list. The remaining criteria (i.e., appearance, enjoyability, and affordability) were ranked in between (Table 3.2).

Table 3. 2. Quality criteria ranked by groups of participants

	Older adults	**Adults with mental health condition**	**Health care Providers**	**App developers**
Ranked as three most important	Security and Privacy	Usefulness	Usefulness	Usefulness
	Ease of use	Cost	Ease of use	Ease of use
	Usefulness	Security and privacy	Security and privacy	Cost
Ranked as middle high	Learnability	Ease of use	Cost	Enjoyability
	Appearance	Learnability	Learnability	Security and Privacy
Ranked as middle low	Cost	Compatibility	Enjoyability	Appearance
	Enjoyability	Enjoyability	Appearance	Learnability
Ranked as three least important	Compatibility	Appearance	Recovery from Error	Recovery from Error
	Recovery from error	Social Acceptability	Social Acceptability	Compatibility
	Social Acceptability	Recovery from Error	Compatibility	Social Acceptability

3.1.1.2.1. Usefulness

Participants in the adults with mental health condition group, app developer group, and health care providers group ranked usefulness as the most important criteria. The older adult participants ranked usefulness as the third most important factor. According to all participants, factors that determined the usefulness of an m-health app were effectiveness, if the app delivered what it promised, and the differences that using an app would make in the users' day to day life. Participants in the health care provider group used terms such as effectiveness, efficacy, content validity, face validity, and functionality to imply usefulness. For convenience, and only in this section, quotes from participants in the adults with mental health condition group are labeled "client".

Older adult 4: *"[to me usefulness is] what is the content and [if] it relates to what I am looking for. … [If the app] gives me the answers that I want. I think that is about it. I mostly use that for information."*

Client 1: *"Whether it is accomplishing the goals that I am setting out to accomplish. Like when I want to track something or express my thoughts if it is doing that thing effectively."*

Health care provider 1: *"Does it do what it says it is going to do? So, I think that is face validity too if you think about it. If I am looking at the title of the app and then I go into the app, is it matching and then does the content also match?"*

App developer 6: *"It is also that it should be functional and do what it is supposed to do. Because if it does not do what it is supposed to be doing, then why would I use it?"*

3.1.1.2.2. Ease of use

Most of the participants ranked ease of use as the second most important factor. Participants said that they expect m-health apps to be self-explanatory, easy to learn and take

very little of their time. Participants in the app developers group used "intuitive" as an alternative term for ease of use. Participants said that ease of use implies factors such as ease of navigation, visibility of the items, simple commands (i.e., use of tap instead of double-tap or pinch), ease of learning, and recovery from errors.

Older adult 5: *"Ease of use of course, once you get in there, can you navigate your way? Are there any problems?"*

Older adult 7: *"[if] I just be able to see clearly what is available in the apps how to get the information quickly, find where the area that I am interested in quickly, and to use it if I have to input and then to get in and out of it sort of to save it, and carry on."*

Client 2: *"Whether I can figure it out or not. [If] there is a logical sequence, things that are close to each other are related, that they are not like back and forth between the screens."*

App developer 5: *"Being intuitive is first, so if I cannot figure it out like in the first two minutes, then I will think that probably there are better apps out there."*

App developer 4: *"Just by like opening the app and like looking at the icons, I should know how to use that without actually having to read anything."*

Some participants believed that ease of use is part of usefulness or contributes to usefulness:

Health care provider 1: *"To me, that [ease of use] goes with usefulness. Because that is part of it and if it is not easy to use then to me, it is useless. For the technology because technology gives the assumption that it is easy, it is meaningful, it is accessible, and if it gets too complex, too many layers, like do this or do that, or there are many menu items that are overwhelming, then I am not going to use it".*

Others believed that ease of use comes from the design of the app.

Client 2: *"It is like internal decorating. For example, things that are related to coffee maker are close to the coffee maker, and you should not walk across the room for that. So, it is a kind of design."*

Some participants suggested including learnability as sub-criteria under ease of use.

Older adult 2: *"well learnability kind of goes with ease of use… [it must be easy] to learn how to maneuver it".*

3.1.1.2.3. Privacy and security

The third most important factor mentioned by the participants was the extent to which m-health apps can protect user's privacy and safeguard the information collected by the app. The main concern among older adult participants were the threats from malware and unauthorized access to their information.

Older adult 2: *"Security of course. Because there is so much malware out there now that is very dangerous."*

Older adult 7: *"I am not comfortable because there are people that they have that, what you call it, when you are typing in your password, they know what keys you are pressing they know!"*

Older adult 5: *"For security, I would ask how easy it is to hack into it."*

Older adult participants named several strategies to protect their privacy and security while using mobile apps. For example, they said that they downloaded apps from a *legitimate* source (by legitimate they referred to iTunes and Google Play store), or asked for help from a trusted friend or relative who had experience with apps. Some used a malware or virus protection software, and others used a secure authentication strategy (e.g., a password or the fingerprint identification feature of the phone).

Older adult 7: *"I am pretty sure that as long as you keep your malware and antivirus apps up to date [you will be safe]."*

Older adult 5: *"You need to have a login."*

Older adult 6: *"I find it quite easy just using my thumbprint. It is more secure than someone guessing your password."*

Older adult 3: *"I would like it to come from a legitimate source; a reliable source so rather than from nowhere because you do have no idea what could happen."*

One of the older adult participants believed that there is very little that older adults can do to protect their security and privacy when they access the internet.

Older adult 1: *"... and the last thing that I put at the very bottom is privacy because of the fact; there is no privacy on the internet… to me, privacy is important, but once it is out there, it is out there, and anybody could get it, and hackers are getting smarter and smarter all the time."*

Some participants expressed a similar sense of helplessness regarding protecting privacy and security in the app developer group:

App developer 1: *"I select the security the last because when you need the app, and you are downloading it from the Google play, your information is already on Google so there is not much you can do about that."*

App developer 6: *"Because nothing is private online. If you have the resources to get the information you want, for example, the credit card information, you can get it."*

The health care provider participants said they are seriously concerned about the privacy of patients and the security of information collected by m-health apps. Health care providers are required to follow strict regulations to protect patients' data. For this reason, they are very

cautious when it comes to recommending m-health apps. In general, health care providers said that they need to know the answers to three questions before they use or recommend an m-health app: 1) what kind of information the app might collect, 2) who may access this information, and 3) what security measures exist in the app to protect patients' data.

Health care provider 1: *"My concern for me is one thing, but for recommending it to clients I am very concerned about their privacy and protecting them from predators really, in terms of marketing and selling false information or products. That is what worries me a lot."*

Health care provider 9: *"As a psychologist, I am always thinking about that. Especially for mental health like where is this information going. If someone is using this information. If they are like using this to track coping [strategies] or symptoms [of clients]. Who is seeing that information? I think people need to be clear about that, so I did lump that with security as well in terms of where that information is going, is it being stored in Canada or even outside of Canada? I would like to have an understanding of that."*

3.1.1.3. Factors perceived as least important

3.1.1.3.1. Social acceptability

Participants in most groups said that social acceptability is not very important when it comes to choosing an m-health app. According to most of the participants, if an intervention or technology worked for someone, it did not matter what other people thought. Besides, the older adult participants believed that society is, in fact, more supportive of seniors using new technologies.

Older adult 1: *"Social acceptability is not really important to me because if it is good for me and is beneficial for me or whoever needs to use it, it does not matter if somebody likes it, it does not matter at all, as long as it is helping the person that needs it."*

Older adult 7: *"I think socially acceptable for us seniors is more likely just to not worry about what other people think; we just say what we want."*

Older adult 5: *"I think [these days] there is more support; it is more acceptable."*

Some of the health care providers, however, said that if the environment (e.g., work, or school) is not supportive of using m-health apps, then it may negatively impact the adherence to the interventions provided through m-health apps:

Health care provider 7: *"For the group that I am working with, a lot of it is return to work. So, we try to pick up apps that they are more likely to use in their workplace and not be embarrassed that to see that their coworkers or their supervisor actually see them use, cause if they are not going to use it at work what is the point of having them practice the skill."*

3.1.1.3.2. Compatibility.

Most users, especially older adults, only own one smart device. Users with one device did not consider the compatibility of the health apps with multiple devices or operating systems as a major concern to them. On the other hand, a few users who owned different devices (e.g., an iPhone and a Samsung tablet) said they preferred to use apps that work on both of their devices. Some of the health care providers said that compatibility is an important factor because no matter how good an app is, they cannot recommend it to their clients who do not use a compatible smartphone or tablet.

Older adult 7: *"Compatibility; again, I don't, I just assume a lot of this thing, I am using the apple iPhone a lot and the tablet, so I pretty well stick with that."*

Health care provider 5: *"I feel like some of my favorite apps that I have used for clients, they aren't available for android, and then I am explaining how to use them in the group*

[therapy session], but I am like, unfortunately, this is only for 'Apple' and like it is really a useful tool, but you cannot use it."

3.1.1.4. Factors ranked in between

3.1.1.4.1. Appearance

For older adults and adults with mental health conditions, size of the fonts and icons, colours, and pattern of the content were important features under the appearance of apps.

Older adults 5: *"I think that appearance is one of the top ones, because if you click on it and you do not like it, or you download it, and you don't like it [you probably won't use it] .... you don't want something that is too busy, cause if it is sort of juvenile looking that will lose my attention."*

Older adults 4: *"Appearance, [if] the colour is good, I find on here that the font is small especially for someone as you get older, your eyesight is not that good, and you may wear glasses."*

Client 5: *"For appearance: colour, fonts, and images that is the main three for me."*

According to the health care providers and app developers, the importance of appearance was for its impact on the ease of use:

Health care provider 5: *"The appearance, I feel like that is, you know, really minimum but in terms of the ease of use I find some of them have too many choices and that could be very overwhelming and when you are anxious or having a panic attack and if you are going through one and you are like I do not know how to use this I cannot figure out."*

App developer 5: *"I also care about the aesthetics, like how clean is the design. Because for some apps they can get really busy, at that point, it might contribute as not as intuitive".*

3.1.1.4.2. Enjoyability

According to participants in the older adults group, the enjoyability of m-health apps was a factor of ease of use and usefulness. Therefore, if an app is easy to use and useful, then users will most probably enjoy using the app. Besides, older adults said that they did not care about using apps as a means of amusement or fun.

Older adult 7: *"If it has to do with health, something like blood pressure, or if I have to put something in or answer something, it [must be] easy to do. I don't care if it is fun or dull. If it gives me the information that I want, I find it enjoyable I don't want to be entertained; I do not need music, if it is information then I am simply in it for the information I do not want to play games."*

Although older adults did not show interest in using fun factors in m-health apps, they said that it was important to them if the app was engaging:

Older adult 1: *"If I download an app and I use it for few times, if I do not like it and find that I am just spinning my wheels, if I do not find it making me more engaged or stimulating my mind, I just delete it, I won't keep it."*

Like the older adults group, adults with mental health condition group valued an app being stimulating and engaging more than it being fun and enjoyable:

Client 5: *"For enjoyability, it has to be stimulating, it has to provide a purposefulness, and it has to have an outcome."*

The app developers used the term "user engagement" to imply enjoyability. While previous scales have suggested time spent on the app as an index for engagement, participants in this focus group suggested the frequency of return to use the app as a better measure to rate engagement:

App developer 5: *"The idea is that if you spend too much time on it, it is more engaging, but with this way, if you like to use it quickly and put it away, then wanting to come back to it, that is where you know if it is engaging or not. For example, you use it and put it away; then in a few minutes or hours, you may want to use it again. So, I guess that may be a better scale for how engaging it is."*

3.1.1.4.3. Affordability

According to the participants in this study, m-health apps should come at a reasonable price, particularly for participants with low income due to disability or retirement. Besides, people with chronic conditions will probably use more than one app. Therefore, if apps are costly or not paid for by grants, health care plans, or insurance policies, many patients with a chronic condition cannot afford them. Most participants suggested $5 as the upper limit for purchasing m-health apps in general however, they also said that they would pay higher amounts if they find the app useful.

Older adult 1: *"It depends, I cannot put a dollar sign depending on what it is; it might be different things. I do not mind spending on things like 'Lumosity' I think I can get it for $70 for two years and I think that is cheap. But so that one is OK but if it were a game or something I would not want to pay $70 because I do not think it is helping my brain as much as the other would be helping my brain so I would rather pay more for my brain than for my enjoyment."*

Client 1*: "Because you need more than [one] app unless you have a group of them. Because I will probably use one app for depression and one for bipolar, they will add up."*

Health care provider 10: *"I feel like it depends, if it is a one-time cost you pay once and you get the app you know forever, or if like you have got a cost coming in every month because then it is harder to justify am I really going to use it now or two months? Am I going to pay this*

money and continue paying for a year or six months and not end up using it and is it really useful at that point especially if you do not like it?"

Health care provider 5: *"Finances are typically a big strain if they are off on disability and even if they spend a couple of dollars on an app and I know it tends to be the better ones you have actually to pay for them or you have to buy the add on so I would try to find both resources, like this costs this much or there is a couple that they are free they do not have as many options or tools on them, but they kind of give you the basics."*

3.1.1.5. Criteria suggested to be added or removed from the framework

Participants suggested several changes to the framework that was presented to them by the investigator. For example, participants in health care provider and adults with mental health condition groups suggested adding a criterion to evaluate if the app or the information provided within the app come from credible sources:

Health care provider 4: *"As this technology is exploding there is the element of which ones are reliable and which ones are sensitive or rigorous enough that one can rely on, there is that quality assurance kind of element because there is a lot out there, and also where is the access point. Where is the source? The Mayo Clinic? Heart and Stroke (Foundation)?"*

Participants suggested that the investigator added *customizability* to the framework. Customizability will evaluate the degree of freedom that users have to customize the features of m-health apps according to their needs and preferences.

Client 1: *"The app should have different categories, such as putting in medications, dealing with friends, colleagues, and work, whatever, sleep, medication food but, that is good but I want to customize it! Yeah customizable is the word."*

Client 7: "*I would like to adjust it also based on my observations or if I changed my mind. In a way that it might change the outcome. Also, customizability of font size and stuff".*

Participants suggested the investigator to use *functionality of the app* instead of *recovery from errors* (i.e., if the app functions with no errors).

Health care provider 7: *"And then recovery from errors, I do not know, I find all technology has errors all the times so just closing it up and being able to restart it I do not think that as a big deal, but if it negatively impacts the overall function of the app then it will be a problem".*

Older adult 2: *"Well this one [The app] wouldn't let me go back with the arrow so I had to go home but it came back up again. It is not the recovery from the error I think, but if there are many errors in it. I think functionality is what I mean".*

Participants in all focus groups suggested that the investigator change "enjoyability" to "user engagement" because they believed m-health apps should be engaging to encourage adherence, and not necessarily fun. They suggested that engagement may depend on how useful the app is, or what features it has that entice users to adhere to using the app.

Older adult 7: *"if it gives me the information that I want, I find it enjoyable, I don't want to be entertained, I do not need music, I do not want to play games".*

Client 5: "*For enjoyability, it has to be stimulating, it has to provide a purposefulness, and it has to have an outcome".*

App developer 5: "*The idea is that if you spend too much time on it, it is more engaging, but with this way if you like to use it quick and put it away, then wanting to come back to it, that is where you know if it is engaging or not. For example, you use it and put it away, then in few*

minutes or hours you may want to use it again. So, I guess that maybe that is a better scale for how engaging it is".

Client 7: "*My problem with them is that I do not stick with them. That is the big issue. Especially when you are trying to keep records of it".*

Health care provider 7: *"if we are recommending it as clinicians for them to use post treatment is it going to be something that they are going to be able to use down the road or is it just going to be like a habit that dies off that you try but it does not engage you maybe that".*

3.1.1.6. User-oriented framework for quality of m-health apps.

The objective of the focus groups was to identify which quality criteria are important from the perspective of m-health app users and app developers. Participants were not able to suggest many criteria on their own. Therefore, the investigator presented them the list of criteria that he had extracted from the usability models and frame works of app evaluation. Participants then discussed which of these criteria were important and ranked them from the most important to the least important. Besides, they suggested the investigator to add, remove, or rephrase some of the criteria he had identified earlier.

The list of criteria presented to the participants included the following: appearance, compatibility, security, cost, ease of use, enjoyability, usefulness, learnability, privacy, social acceptability, and recovery from errors. Participants in the focus group ranked usefulness, ease of use, privacy, and security as the most important criteria. In contrast, they ranked social acceptability and compatibility as the least important criteria and suggested that these criteria should be excluded from the framework. In addition, they suggested that learnability could be covered under ease of use. They also suggested the investigator add trustworthiness and customizability to the framework. Finally, they recommended the investigator to rename cost,

enjoyability and recovery from errors, to affordability, engagement, and functionality, respectively.

Next, the investigator revised the initial list of quality criteria that he had extracted from the usability models and frameworks of app evaluation and developed the user-oriented version of the framework for evaluating mobile health apps (Figure 3.1).

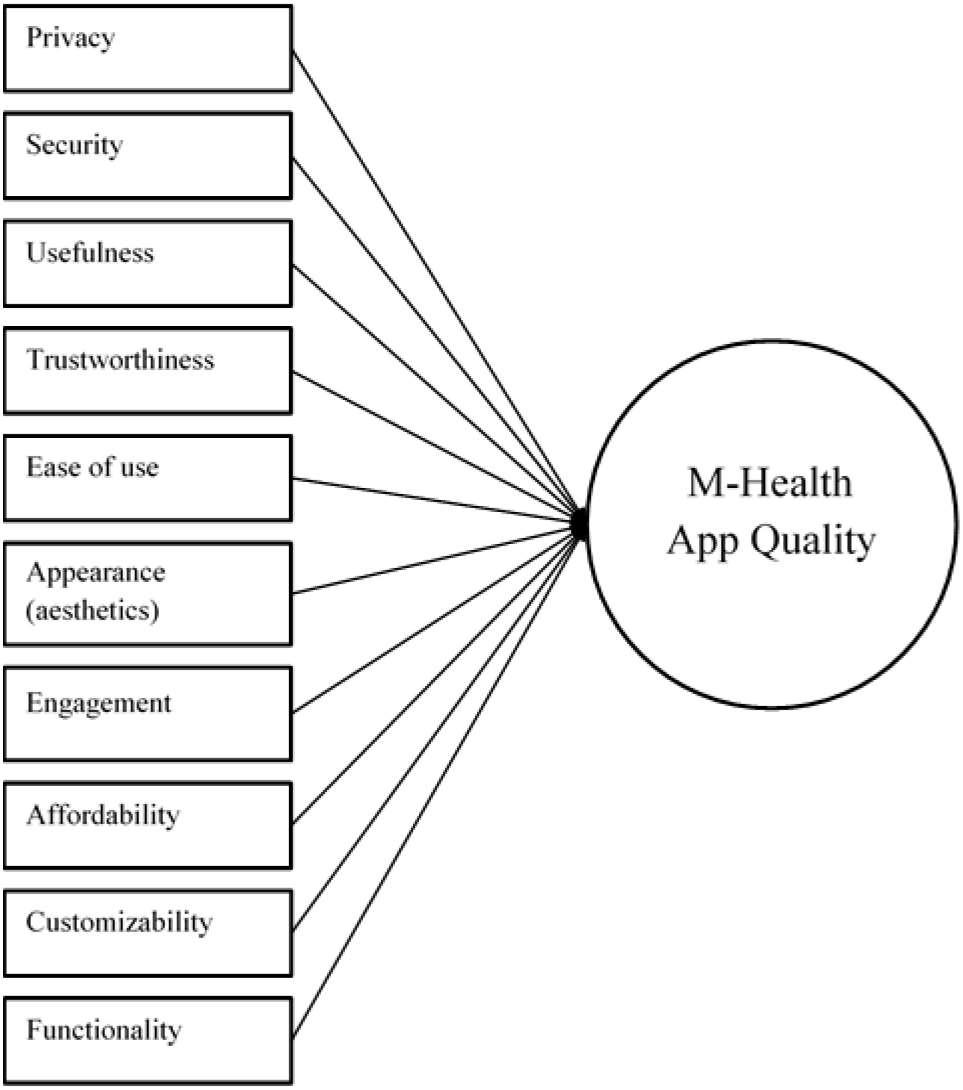

Figure 3. 1. User-oriented framework for quality of m-health apps

3.1.1.7. App quality: A composite construct

Most often, researchers are interested in measuring hypothetical constructs that are not directly measurable. However, it is possible to measure the observable manifestations of such constructs using items of a scale. According to the *Classical Test Theory* (CTT), a universe of items exist that tap into the hypothetical construct. Since the same underlying construct has caused them all, they are referred to as *effect indicators* (they are the effects of the hypothetical construct). Also, the items in the scale must be homogenous. In other words, they must correlate with each other.

For example, we may not be able to measure anxiety directly. However, from research, we may know that anxiety causes symptoms such as sweatiness, worrying, and sleep disturbances. If we develop items to measure such manifestations, we may indirectly measure anxiety.

However, there are several situations that none of these apply. Some constructs do not cause effects in the environment. Instead, few criteria (causal indicators) define them. These constructs are referred to as *composite constructs* [101]. Casual indicators are independent of one another. Therefore they are not correlated with each other. As a result, the assumptions of homogeneity from CTT do not apply in such cases. The assessment tools comprised of causal indicators to measure composite constructs are usually referred to as *indices*.

An example is measuring recent stressful life events that may include items as the death of a spouse, receiving a driving ticket, or purchasing a new house [101]. Logically neither of these are correlated to the other. However, all of them may have a role in *causing* a stressful life.

According to the user-oriented framework developed in phase 1, one may conclude that app quality is most probably a composite construct defined by at least 10 quality criteria

identified in phase 1. As a result, to measure the quality of mobile health apps, it is advised to follow guidelines for developing an index rather than a scale, the practical implications of which are as follows:

- Items do not have to correlate to each other
- Assumptions of homogeneity and relevant statistical procedures (i.e., internal consistency, Cronbach alpha, and factor analysis) do not apply here
- The index must cover all quality criteria that measure the quality of mobile health apps

3.1.2. Generating the Item Pool

In the next step, the investigator generated an item pool for each of the following quality criteria identified in the previous study: privacy, security, usefulness, trustworthiness, ease of use, appearance (aesthetics), engagement, affordability, customizability, and functionality.

The initial plan of the investigator was to generate the item pool based on the suggestions made by the participants in the focus groups in the previous phase. However, none of the participants were able to suggest items to assess the identified quality criteria of mobile health apps. As an alternative, the investigator used the scales developed in previous studies as a source to generate the item pool for the index.

The investigator conducted a systematic review of the literature following *PRISMA* statement (Preferred Reporting Items for Systematic reviews and Meta-Analyses) [110] to identify scales and questionnaires used in previous studies to rate the quality and usability of apps. Next, he extracted the items from the identified scales and mapped them to the ten quality criteria that were identified in the previous study.

3.1.2.1. Characteristics of identified sources for scale items

In total, 4402 papers were identified. After removing 725 duplicate records, the investigator and a research colleague (Noelannah Neubauer) reviewed the title and abstract of 3677 remaining papers, from which 206 papers met the inclusion criteria and were selected for full-text review. After reviewing the full text, 122 additional papers that did not meet the inclusion criteria were excluded. They also reviewed the references of the papers eligible for full-text review and identified three additional articles that met the inclusion criteria for data abstraction. Finally, they selected 87 papers for data extraction. The investigator and his colleague did not find any additional scales or questionnaires in the grey literature. The review process is presented in Figure 3.2.

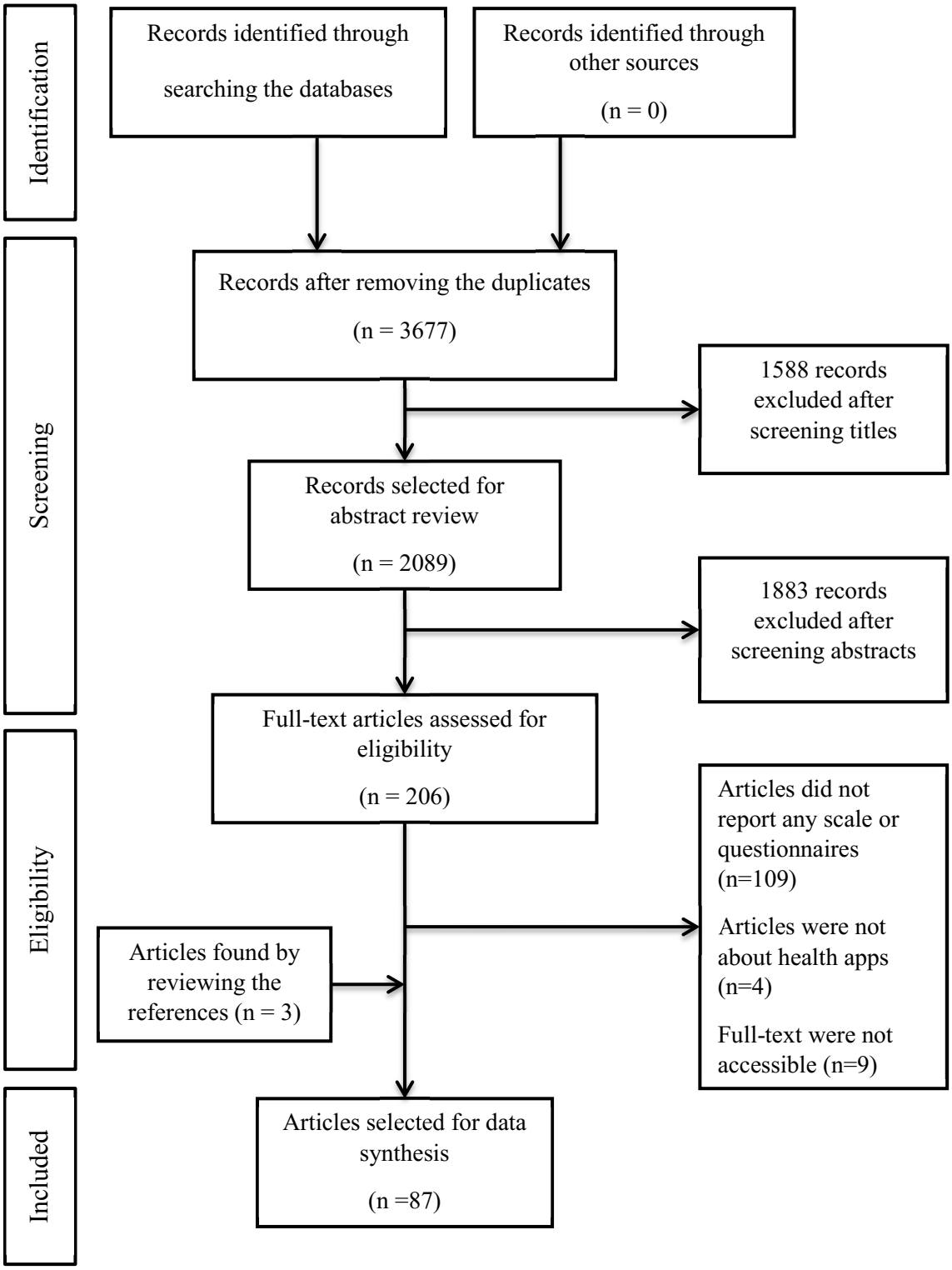

Figure 3. 2. Review process

3.1.2.2. Scales and questionnaires

A total of 48 rating scales for health apps were identified in 87 studies included in this review. Of these scales, 23 were app usability scales, and 25 were app quality rating scales (Table 3.3).

Usability refers to a multi-dimensional property of a computer interface that is associated with attributes such as ease of use, learnability, efficiency, memorability, error rates and user satisfaction [33]. However, such scales do not cover criteria that are important to health care providers such as security of the app, customizability of the features, appropriateness to the users, and trustworthiness of the health information provided [60]. Therefore, attempts have been made to create quality rating scales for m-health apps to cover the criteria that are important to health care providers as well as usability experts.

Most of the usability scales that we identified in this review were scales initially used for usability assessment of other information technologies such as computers or websites. Examples are *Perceived usefulness and ease of use questionnaire (PUEU)* [111], *Software Usability Measurement Inventory (SUMI)* [42], *Computer system usability questionnaire* [112], and *System Usability Scale (SUS)* [34]. On the other hand, scales such as the *ISOMETRIC* [113], *Usefulness, Satisfaction, and Ease of use (USE) Questionnaire* [114], *Health Information Technology Usability Evaluation Scale (Health-ITUES)* [115], and *Quality of Experience (QOE)* [116] were exclusively developed for usability assessment of mobile health apps. The investigators also identified 15 studies that used a different scale to evaluate usability of m-health apps, however, none of these studies had named the usability scale they used or developed (Table 3.3).

In addition to the usability scales, the review identified 25 scales that rated criteria which were not covered by the usability scales. These scales were generally named as quality rating

scales. However, unlike the usability scales, none of the quality rating scales were based on a clear definition of the quality of mobile health apps. In addition, none of the quality rating scales were developed based on any of the frameworks for evaluating the quality of m-health apps. Some of the quality rating scales were exclusively developed to rate the quality of apps that were relevant to specific health condition. These scales were *APPLICATION* [61] (for pregnancy wheel apps), *Enlight* [64] and *NICE* [117] (for change of behavior apps), *Fitness app scoring instrument* [118] (for fitness apps), *MedAd-AppQ* [66] (for medical adherence apps), *Nutrition App Quality Evaluation (AQEL)* [65] (for nutrition apps), and *Quality Assessment tool for Evaluating Medical Apps (QAEM)* [63] (for medication complication apps).

The investigator identified 18 scales that were developed to rate the quality of m-health apps in general: *App Chronic Disease Checklist* [67], *Health Care Apps Evaluation Tool* [68], *Royal College of Physicians' Health Informatics Unit checklist* [69], *Organisation for the Review of Care and Health Applications- 24 Question Assessment (ORCHA-24)* [70], and *Mobile Application Rating Scale (MARS)* [5, 71-86] and its users version *U-MARS* [87]. The investigator identified one study that used three different scales to evaluate the quality of panic disorder apps: *Silberg Scale* (for accountability of the information provided), *Discern scale* (for transparency and quality of information) and *Abbott's scale* (for interactivity) [119]. Besides, the same study and three others used recommendations of Health on the Net Foundation (HON-code) [99] to rate the quality of the information provided in different health apps [119-122]. Finally, ten studies reported developing an app quality rating scale but did not name the scale that they developed (Table 3.3).

Table 3. 3. Scales identified in the literature

App Usability Scales [references]	App Quality Scale [references]
Perceived Useful and Ease of use questionnaire (PUEU) [41]	APPLICATION scoring system [61, 62]
Software Usability Measurement Inventory (SUMI) [42]	Quality Assessment tool for Evaluating Medical Apps (QAEM) [63]
System Usability Scale (SUS) [35, 38, 39, 43-52]	NICE guidelines [117]
ISOMETRIC [53]	Enlight [64]
USE Questionnaire [54, 55]	Fitness app scoring instrument [118]
Quality of Experience (QOE) [116, 123]	Nutrition App Quality Evaluation (AQEL) [65]
Health-ITUES [56-58]	MedAd-AppQ [66]
Computer System Usability Questionnaire [59]	Silberg Scale [120, 124-126]
Not named [127]	Brief Discern scale [119]
Not named [128, 129]	The modified version of Abbott's scale [119]
Not named [130]	App Chronic Disease Checklist [67]
Not named [131]	Health Care Apps Evaluation Tool [68]
Not named [132]	Royal College of Physicians' Health Informatics Unit checklist [69]
Not named [133]	The Organisation for the Review of Care and Health Applications- 24 Question Assessment (ORCHA-24) [70]
Not named [134]	Mobile Application Rating Scale (MARS) [5, 71-86]
Not named [135]	Mobile Application Rating Scale-User version (U-MARS) [87]
Not named [136]	HON-Code [119-121]
Not named [137]	Not named [138-140]
Not named [141]	Not named [142]
Not named [143]	Not named [144]
Not named [145]	Not named [146]
Not named [147]	Not named [148]
Not named [149]	Not named [150]
	Not named [151]
	Not named [152]

3.1.2.3. Classification of scale items

The investigator extracted 771 items from the scales identified in the review. The investigator grouped similar items under the same quality criteria. In the next step, a panel of

experts with expertise in the field of technology acceptance and scale development (Adriana Rios Rincon, Antonio Miguel-Cruz, and Mary Roduta Roberts) independently reviewed the items mapped to each criterion and rated the extent to which each item aligned with the suggested criterion using a 3-point Likert scale (1= item does not align at all, 2= item aligns moderately with suggested construct, and 3= item aligns perfectly).The investigator found the median of the ratings by all reviewers. Items with a median rating score of 3 across all three raters (i.e., at least 2 raters rated the item as aligns perfectly) were considered as *correctly mapped*. The reviewers agreed that 82% (632/771) of items were mapped correctly by the investigator (Table 3.4).

Table 3. 4. Summary of the initial mapping of the scale items

Criteria	**Number of items mapped to each criterion by PA**	**Number of items reviewers agreed were correctly mapped (%)**
Aesthetics	52	31 (56)
Affordability	3	3 (100)
Customizability	14	13 (93)
Ease of use	174	160 (92)
Functionality	27	22 (81)
Privacy	40	31 (77.5)
Satisfaction	19	19 (100)
Security	29	24 (82.75)
Trustworthiness	112	98 (87.5)
Usefulness	204	175 (86)
User Engagement	13	9 (69)
Not categorized	84	47 (56)
Total	771	632 (82)

In the next step, the investigator revised the classification of items based on the reviewers' feedback (Table 3.5). In total, reviewers suggested that 18% of the items (139/778) to be classified under a different criterion. The investigator placed these items under the criteria suggested by the majority of the reviewers (2 out of 3). In cases where reviewers did not agree with each other, the criteria suggested by Antonio Miguel-Cruz was accepted since he had more experience with theories of usability and acceptance.

Moreover, 8% of the items (69/771) did not match any of our criteria. These items were either too specific to certain apps or health care conditions (e.g., "*Definition of enuresis [in the app] is in accordance with ICCS"* [142]) or they were not relevant to measuring quality of apps (e.g., "*Adequate sample sizes for evaluation of app ensured"* [117]) and therefore were not used in the item pool (Appendix J).

Table 3. 5. Final distribution of items under the quality criteria

Criteria	**Number of Items under each criterion**
Aesthetics	37
Affordability	3
Customizability	21
Ease of use	195
Functionality	31
Privacy	34
Satisfaction	33
Security	30
Trustworthiness	107
Usefulness	185
User engagement	26
Not categorized	69
Total	771

The investigator used the remaining items to generate an item pool with 74 items distributed unevenly under ten quality criteria (Appendix C). Under each category, the investigator merged items that measured the same attribute. For example, the item "*Is the privacy policy available within the app*" [45]; was merged with a similarly worded item to *"Is there a data privacy policy, either within the app itself or on a website"* [71]. The final item pool is available in Appendix C.

3.2. Phase Two – Developing ARIA

The objective of this phase was to develop the first draft of ARIA. To achieve this objective the investigator conducted two research activities. In the first activity, the investigator assessed the relevance and representativeness of items in the item pool. In the second activity, the investigator recruited a panel of experts to short list the items that were identified as relevant in the first activity and develop the first draft of ARIA.

To facilitate the content validation process in the first research activity, the investigator developed an online survey using the SurveyGizmo® platform and e-mailed the link to a group of potential users of m-health apps including health care providers, patients with lived experience of mental health conditions, and older adults. Appendix C presents a paper-based version of the survey.

3.2.1. Participants' Characteristics

Eighteen participants completed the survey. Participants were from four provinces of Canada (Alberta, Ontario, Quebec, and Newfoundland & Labrador). Half of the participants were below 50 years of age. Participants were from different stakeholder groups, including health care providers (8), older adults (1), adults with mental health condition (6), and app

developers (3). All participants owned and used a smartphone for at least four years, and 13 of them also owned a tablet (Table 3.6).

Table 3. 6. Characteristics of survey participants

Participant group	Gender (N) Female/Male	Age range (N)	Province of Residence
Health care provider	7/1	30 – 39 (3) 40 – 49 (1) 50 – 59 (2) 60 & above (2)	Alberta: 7 Newfoundland & Labrador: 1
App developer	2/1	30 – 39 (2) 40 – 49 (1)	Ontario: 2 Quebec: 1
Adults with mental health condition	1/5	20 – 29 (1) 30 – 39 (1) 50 – 59 (2) 60 & above (2)	Alberta: 5 Ontario: 1
Older adults	0/1	60 & above (1)	Alberta: 1

3.2.2. Content validation of the item pool

To quantify the level of agreement of participants with the relevance of the items to the given criteria, the investigator calculated the content validity index (CVI) for each item. The CVI is the proportion of survey participants that rate an item as relevant to a given quality criterion. Following Lynn's recommendation [105], items endorsed by at least 80% of the survey participants as relevant (i.e., CVI $\geq$ 0.8) were considered valid. In total, 47 items met this criterion (Table 3.7).

Table 3. 7. Summary of the content validation survey results

Criteria	Total number of items	Number of Items with CVI ≥0.8	CVI Average (Range)
Aesthetics	5	3	0.83 (0.67 – 0.94)
Affordability	4	2	0.85 (0.78 – 1.00)
Customizability	5	5	0.86 (0.83 – 0.89)
Ease of use	15	11	0.85 (0.61 – 1.00)
Functionality	6	4	0.86 (0.61 – 1.00)
Privacy	6	1	0.74 (0.61 – 1.00)
Security	7	7	0.91 (0.83 – 0.94)
Trustworthiness	13	8	0.79 (0.5 – 0.89)
Usefulness	7	4	0.83 (0.67 – 1.00)
User engagement	6	2	0.75 (0.56 – 0.89)
Total	74	47	0.83 (0.5 – 1.00)

A complete list of items and their associated CVIs is available in Appendix K. The investigator deleted the items with CVI below 0.8, except for three items. Under Privacy, the investigator retained item number five (the item that evaluated if an app accessed the phones sensors such as microphone and camera) since this seemed to be a major concern from the interviews in phase 1 and also the CVI for this item was 0.77, which was very close to 0.8 threshold. Under appearance and aesthetics, the investigator decided to keep item number 49 because it had been used in 12 other scales identified in the literature, and its CVI was 0.78. Under Affordability, the investigator decided to keep item number 66 since it was important to users according to the focus groups in phase 1 and also the CVI for this item was also very close to the threshold (i.e., 0.78).

Moreover, the investigator made a number of revisions to the items as suggested in the comments made by the survey participants. Under privacy, the investigator broke item number one into four items to facilitate answering this item. Under Security, the investigator combined item 11 and 12 because they were very similar. Under user engagement, the investigator added one item to the pool (i.e., *the app uses various ways such as notifications, reminders, or achievement points to engage users*) based on a suggestion made by one of the participants with app development background. As a result of these changes, the total number of items in the revised version of the item pool increased to 53 (Appendix D).

3.2.3. Shortlisting the validated items

In the first activity of phase two, a quantitative method (CVI) was used to identify the relevant items to each criterion. The objective of the second activity in phase two was to use a qualitative method to determine the optimum length of the index by removing redundant or items

that were not very important while maintaining the content coverage of the index (i.e., to ensure that the index covers all relevant quality criteria).

To achieve this objective, the investigator used the revised version of the item pool (developed in 3.2.2) and created an item review form (Appendix D). The form was sent to a panel of m-health app consumers including two older adults, one adult with a mental health condition, one expert in usability evaluation, one occupational therapist, and one psychologist. The panel members were asked to review the items on the form and recommend which ones to keep and which ones to remove to reduce the length of the index while maintaining its content coverage. The panel members had one week for this task.

Next, the panel members participated in a focus group. During this focus group, the panel members discussed and agreed by consensus on which items should be kept on the index and which ones should be removed. In addition to the panel members, the investigator and two members of his supervisory committee (Drs. Lili Liu and Mary Roduta Roberts) were present during the focus group as facilitators. The facilitators did not participate in the voting process. Table 3.8 presents the summary of the decisions and votes made by the panel members.

Table 3. 8. Summary of the decisions made by the panel members

Item number	Voted to keep the item (n)	Voted to remove the item (n)	Final recommendation (after discussions)	New Item number (Appendix E)
1	3	3	Merge 1-5	5
2	4	2	Merge 1-5	5
3	4	2	Merge 1-5	5
4	4	2	Merge 1-5	5
5	2	5	Merge 1-5	5
6	4	2	Revise	7
7	2	4	Delete	NA
8	2	4	Delete	NA
9	5	1	Delete	NA
10	4	2	Revise and merge with 6	7
11	4	2	Revise and merge with 6	7
12	4	2	Revise	2
13	4	2	Revise	9
14	4	2	Keep	4
15	5	1	Revise and merge with 12	2
16	4	2	Delete	NA
17	3	3	Revise and merge with 14	4
18	3	3	Revise	3

Table 3. 8. Continued

Item number	Voted to keep the item (n)	Voted to remove the item (n)	Final recommendation (after discussions)	New Item (Appendix E)
19	4	2	Revise and merge with18	3
20	4	2	Revise	10
21	4	2	Delete	NA
22	4	2	revise	11
23	3	3	Revise and merge with 22	11
24	5	1	Delete	NA
25	6	0	Revise and merge with 22	11
26	6	0	Revise	12
27	4	2	Revise and merge with 22	11
28	5	1	Revise and merge with 22	11
29	5	1	Delete	NA
30	3	3	Revise merge with 22	11
31	5	1	Delete	NA
32	5	1	Delete	NA
33	6	0	Delete	NA
34	3	3	Delete	NA
35	5	1	Revise	15
36	4	2	Delete	NA

Table 3. 8. Continued

Item number	Voted to keep the item (n)	Voted to remove the item (n)	Final recommendation (after discussions)	New Item (Appendix E)
37	3	3	Delete	NA
38	3	3	Revise	14
39	5	1	Revise and merge with 38	14
40	5	1	Revise and merge with 38	14
41	4	2	Revise and merge with 38	14
42	4	2	Revise and merge with 38	14
43	6	0	Delete	NA
44	5	1	Delete	NA
45	5	1	Keep	13
46	4	2	Delete	NA
47	6	0	Revise	6
48	3	3	Revise and merge with 47	6
49	2	4	Revise and merge with 47	6
50	6	0	Revise	1
51	4	2	Delete	NA
52	4	2	Revise merge with 50	1
53	2	4	Delete	NA

3.2.3.1. Focus group findings

Most of the panel members voted for fewer items to make the scale easy to use for the general public. They said that many of the items were redundant and could be merged with other items or deleted without reducing the content coverage of the index. At the end of the session, the panel members made the following recommendations.

3.2.3.1.1. Format of the index

The panel members recommended that the index should have two parts. The first part of the index should include items that pertain to screening an m-health app before downloading it on one's mobile device. Examples are items that rate the purpose of the app, trustworthiness of the developers, and affordability of the app based on the information available on the app store or the app developers' website. The second part of the index should include items that rate the usability and functionality of m-health apps.

3.2.3.1.2. Privacy and Security

Panel members recommended that the index has at least one item on the privacy policy of the app. They also recommended that the privacy policy in the m-health app must address at least three key areas: (1) what information will be collected by the app, (2) who will have access to this information, and (3) how this information will be used. Moreover, they recommended that at least one item be included in the index to rate the ways the app protects users' information from unauthorized access.

3.2.3.1.3. Trustworthiness

Panel members recommended that at least one item be included to address each of the following factors relevant to trustworthiness of the app: the credibility of the developers, risk assessment, conflict of interest, and source of information.

3.2.3.1.4. Ease of use

Regarding the ease of use, the focus group recommended merging items from other criteria such as functionality and aesthetics under ease of use. They suggested that the new construct include the following items: ease of navigation, visibility of the components, comprehensibility of the information, functionality, and that the app is error-free.

3.2.3.1.5. User engagement

Participants suggested that the investigator remove this criterion. They justified that user engagement is a factor of ease of use, functionality, and visual appeal. According to the participants, since these items have already been covered in the previous criteria, a standalone criterion for engagement would be redundant. However, participants suggested that two items be added to measure general satisfaction and to assess if the app is pleasing to use.

3.2.3.1.6. Customizability

Participants suggested one item to rate general customizability of app settings and functions such as font size, language, sound, and notifications.

3.2.3.1.7. Affordability

Participants recommended one item to rate the degree to which users find the app affordable. They suggested adding cues such as being aware of download fees, subscription fees, and any free versions.

3.2.3.1.8. Usefulness

Participants suggested using two items for this criterion. For part A, they suggested an item that rates the extent to which the purpose and goals described in the app are relevant to the user. For part B, they suggested one item to rate the degree to which users find apps useful (i.e., perceived usefulness).

3.2.3.1.9. Appropriateness of the app for the target group

Participants suggested that the scale needs one item to rate the extent to which the content of an m-health app is appropriate for the users considering their age, gender, culture, and education. This factor was also mentioned in the framework for the assessment of mental health apps by the Mental Health Commission of Canada (MHCC). In the original item pool, this item was under Aesthetics. However, because of its importance as pointed out by the focus group participants, and to comply with the MHCC framework, the investigator created a new criterion with one item to address the appropriateness of the content.

3.2.3.2. Developing the pilot draft of ARIA

Based on the recommendations of the panel members, the pilot draft of ARIA was developed (Appendix E). In addition to items that were generated from the item pool, four new items were generated for the pilot version based on recommendations made by the panel members (items 8, 16, 17, and 18).

3.3. Phase Three – Testing ARIA

This phase consisted of two activities. First, the first draft (pilot) of ARIA was pilot tested by a small group of potential target users of m-health apps. Next, the investigator revised the pilot draft based on the comments and recommendations made by the participants in the pilot study and created the final version of ARIA. In the second research activity, 12 potential users of m-health apps, used the final version of ARIA to rate the quality of 11 m-health apps in the field of mental health. The investigator used these ratings to assess the inter-rater reliability of ARIA in a generalizability study (G-study).

3.3.1. Pilot testing

Two older adults and seven adults with a mental health condition used the pilot draft of ARIA to rate the quality of one m-health mobile app in the investigator's presence (Table 3.9). Participants were requested to *think aloud* and describe their thought process while applying the index to the given app. Moreover, they suggested any changes or revisions that they thought were needed to improve the readability of the index items.

Table 3. 9. Participants in the pilot study

Participant group (N)	Mean Age (SD)	Gender (F/M)	Phone used (iPhone/Android)
Older adults (2)	73 (7)	0/2	2/0
Adults with mental health condition (7)	53 (8.9)	2/5	3/4

3.3.1.1. Findings from the pilot study

The findings from the interviews during the pilot study are as follows.

3.3.1.1.1. Instructions

Participants found the instructions for part B easier to understand than the instructions for part A. Many of the participants were unable to find the answers to the items in part A. However, after the investigator provided adequate cues, all participants were able to answer the questions in part A. They recommended the investigator to revise the instructions in part A by providing details on where to find the information to answer each item in this section (see Appendix F and G, added instructions are in bold face letters). Moreover, almost all the participants forgot to calculate the overall score for each part. They mentioned that it was not implied that they needed to add up the scores, and this direction should be added at the end of each part of the index.

3.3.1.1.2. Scaling anchors

Participants found the scaling anchors (i.e., not at all true – completely true) difficult to understand. Some selected *not at all true* where they should have selected *not applicable*. When prompted by the investigator, they said that in their opinion, when a statement was not true; hence, it was not applicable. They recommended using the traditional 5-point Likert scales (*completely disagree – completely agree* with *neutral* as the midpoint) and removal of *Not applicable* option to reduce risk of confusion.

3.3.1.1.3. Index constructs

Participants recommended the investigator to provide the quality criteria in a sidebar on the left side of the questions. They believed providing the criteria would help users better understand and interpret the ratings.

3.3.1.1.4. Overall star-rating system

Participants recommended that the investigator add an overall star rating for the quality of apps at the end of the index. They justified that this will help users to provide their overall rating of the quality of the app using a visual tool.

3.3.1.2. Developing the final version of ARIA

The investigator revised the index based on the recommendations received from the participants in the think aloud study. To improve the comprehensibility of the items, participants recommended the investigator to change the tone of the questions from the third person (e.g., "the content of the app is appropriate for the intended users of the app") to the first person (e.g., "the content of the app is appropriate for me"). While this would improve the comprehensibility of the index items for patients, it may make the items awkward for care providers who usually rate the quality of the apps on behalf of a third person. Therefore, the investigator created two

versions of the index: one for health care providers (Appendix F), and one for the general users of m-health apps (Appendix G).

3.3.2. Reliability testing

In this phase, 12 potential users of m-health apps (Table 3.10) used ARIA to rate the quality of a sample of 11 apps. The objective of this study was to investigate the extent and sources of variability of total scores of ARIA when it was applied to assess a sample of mental health mobile apps by different app users.

Table 3. 10. Raters' characteristics

Participant group (N)	Gender (F/M)	Mean Age (SD)	Phone used (iPhone/Android)
Occupational therapists (4)	4/0	35 (8.3)	2/2
Older Adults (4)	0/4	66.5 (8.5)	2/2
Adults with a mental health condition (4)	2/2	49 (9.3)	2/2

Three groups of raters were recruited: four health care providers, four older adults, and four adults living with a mental health condition. The investigator instructed the participants in the health care provider group to rate the apps based on a detailed mock patient profile (Appendix I). Participants in the older adults and adults with a mental health condition groups were instructed to rate the apps according to their needs and expectations (i.e., as if they wanted to use the apps themselves). For this reason, the results were mainly analyzed by groups collapsed across raters.

3.3.2.1. Distribution of app total scores by raters within each group

Figures 3.3 – 3.5 presents the distribution of scores for each app by raters in each group. The total possible scores for the index ranged between 0 and 72 points. The lowest score was 19, given to the *Headspace* app by rater 1 in the older adults group (Figure 3.4). The highest score

was 66, given to *Moodjournal* app by rater 4 in the adult with mental health condition group (Figure 3.5).

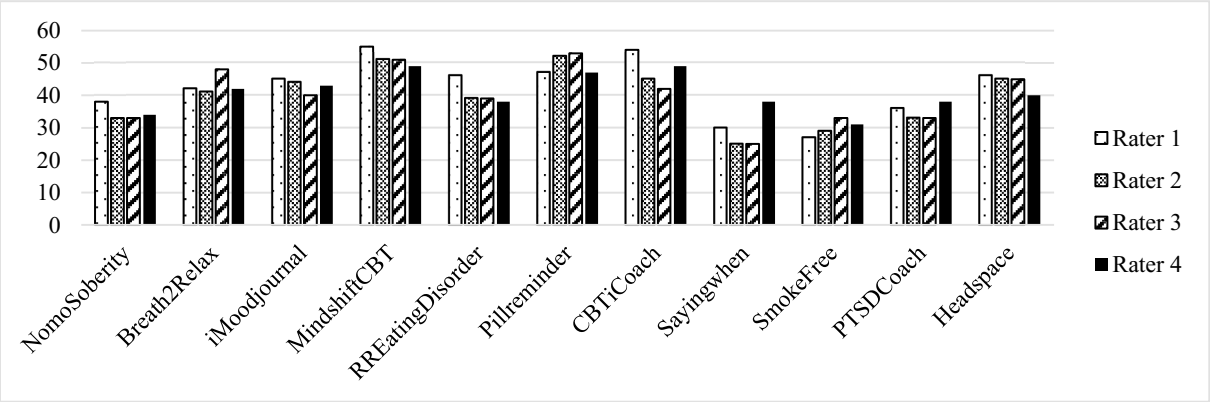

Figure 3. 3. Distribution of total scores of ARIA for each app by raters in health care provider group

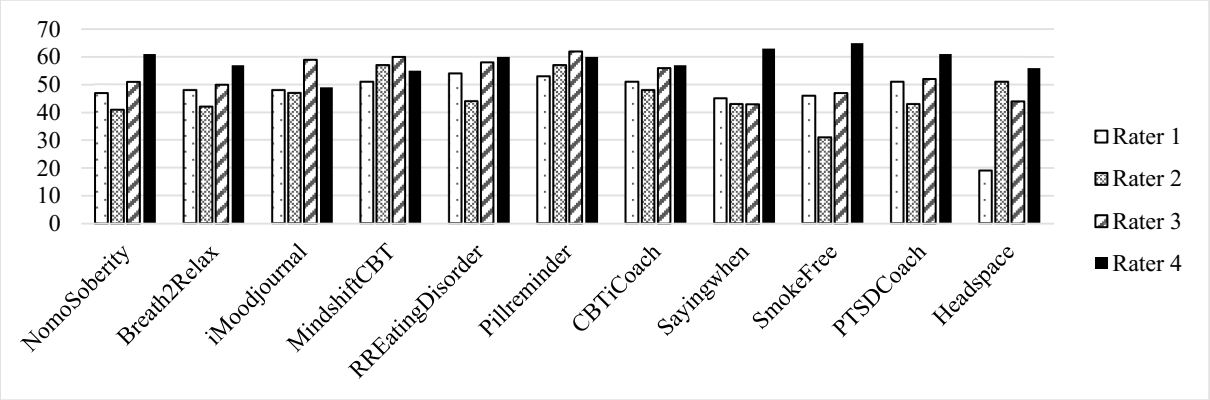

Figure 3. 4. Distribution of total scores of ARIA for each app by raters in older adults group

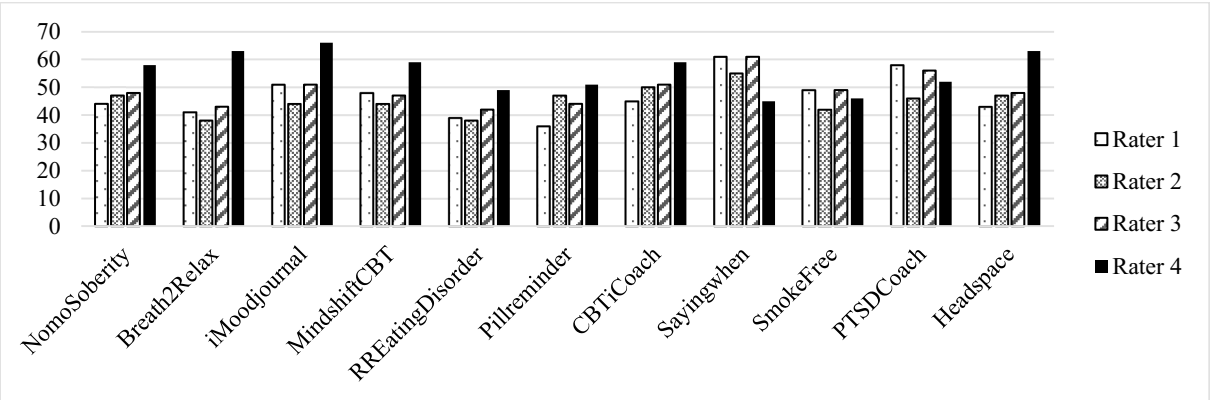

Figure 3. 5. Distribution of total scores by raters in adults with mental health condition group

Of all scores given by all raters, only 2 (1.5%) were within 10% of the best possible scores for the index (i.e., between 64.8 and 72). None of the scores were within the 10% of worst possible scores for this index (i.e., between 0 and 7.2). The distribution of the total scores does not signal a potential ceiling or floor effect for total scores of ARIA.

3.3.2.2. Comparing average total scores across groups and raters

The range and distribution of total scores for each app varied between different groups of raters (Figure 3.6). For example, occupational therapists rated *SayingWhen* (an app for managing drinking problem) the lowest (29.5, SD=6.13) and *Mindshift-CBT* the highest (51.5, SD=2.5). On the other hand, the older adults rated *Headspace* (an app to teach mindfulness) the lowest (42.5/72, SD=16.42) and the *Pillreminder* (an app to help with medical adherence) the highest (58/72, SD=3.91). The adults with a mental health condition group rated the *RREatingDisorder* (an app to help with eating disorders) the lowest (42/72, SD=4.96) and the *Sayingwhen* the highest (55.5/72, SD=7.54).

The observed between-group differences suggest that raters in each group have different needs and preferences, which indicates that ARIA may differentiate between the needs, goals and expectations of the app users. For example, the mock patient profile given to the OTs stated that the patient is not a smoker. Therefore, the *SmokeFree* app, for example, was irrelevant to this patient. As a result, the occupational therapist participants rated the *SmokeFree* app much lower than the other participants who probably found it more relevant to them.

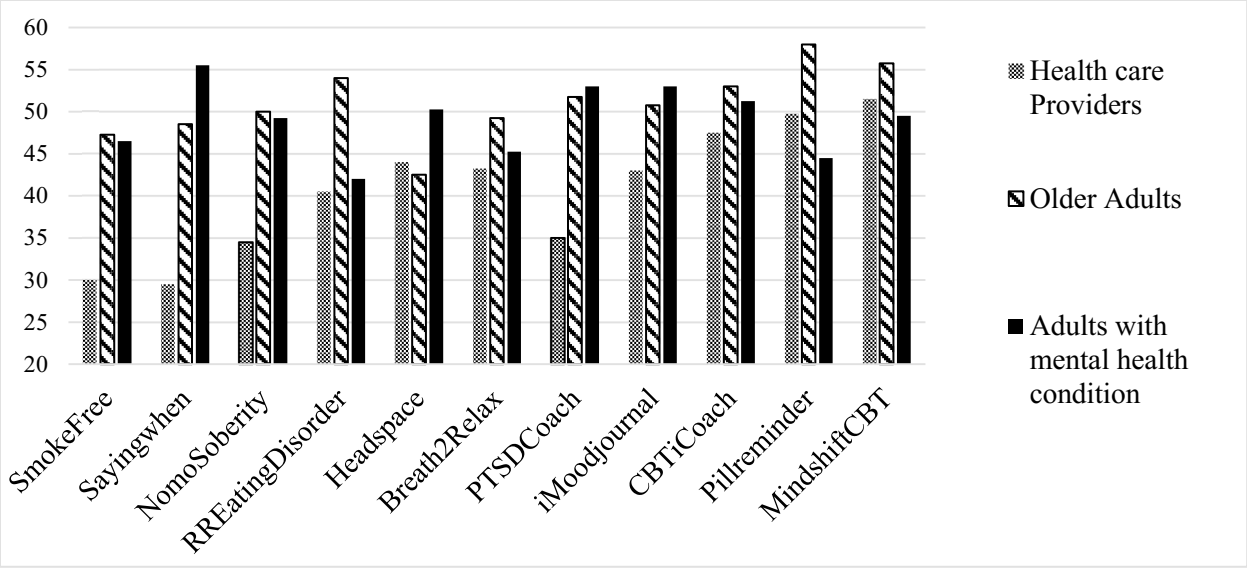

Figure 3. 6. Distribution of total score of ARIA for apps by each group

Raters in each group combined

The investigator conducted 3 x 4 (Group x Rater) mixed factorial analysis of variance (ANOVA) to investigate the following:

1) If the average total score given by health care providers, older adults, and adults with mental health condition are significantly different from each other (i.e., between subject factor with three levels: health care providers, older adults, and adults with mental health condition).

2) If raters within each group, have rated the apps significantly different from each other (i.e., within subject factor with four levels of raters within each group)

Mixed factorial Analysis of variances (ANOVA) confirmed a significant between subject (Group) effect ($F(2, 30) = 10.45$, $p<0.01$) with the health care provider group rating the apps significantly lower than the older adults group (mean difference = 10.2, 95% CIs [16.2, 4.182]) and the patient group (mean difference =8.31, 95% CI [14.34, 2.29]) (Figure 3.7).

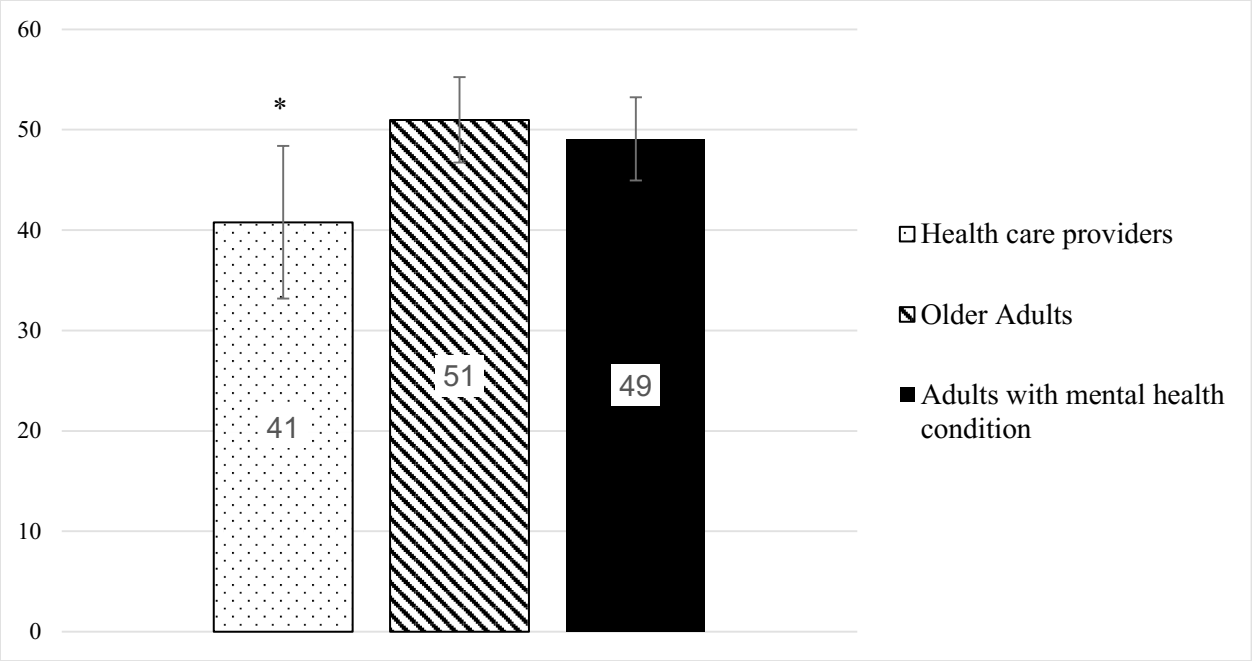

Figure 3. 7. Average total scores of ARIA for all apps by each group of raters.

*Significant difference with other groups at P<0.001

The mixed factorial ANOVA indicated a significant two-way (Group x Rater) interaction effect ($F(6, 90) = 3.8$, $p<0.05$). Therefore, simple effect analysis of the total scores within each group of raters was conducted, using Bonferroni correction to identify if raters within each group scored the apps significantly different from each other. The results indicated no significant differences in the total scores within the health care provider group. However, the total scores by two participants in the older adults group (raters 3 and 4) and the total score of one participant in the adults with mental health condition group (rater 4) were significantly higher in comparison to their peers (Figure 3.8).

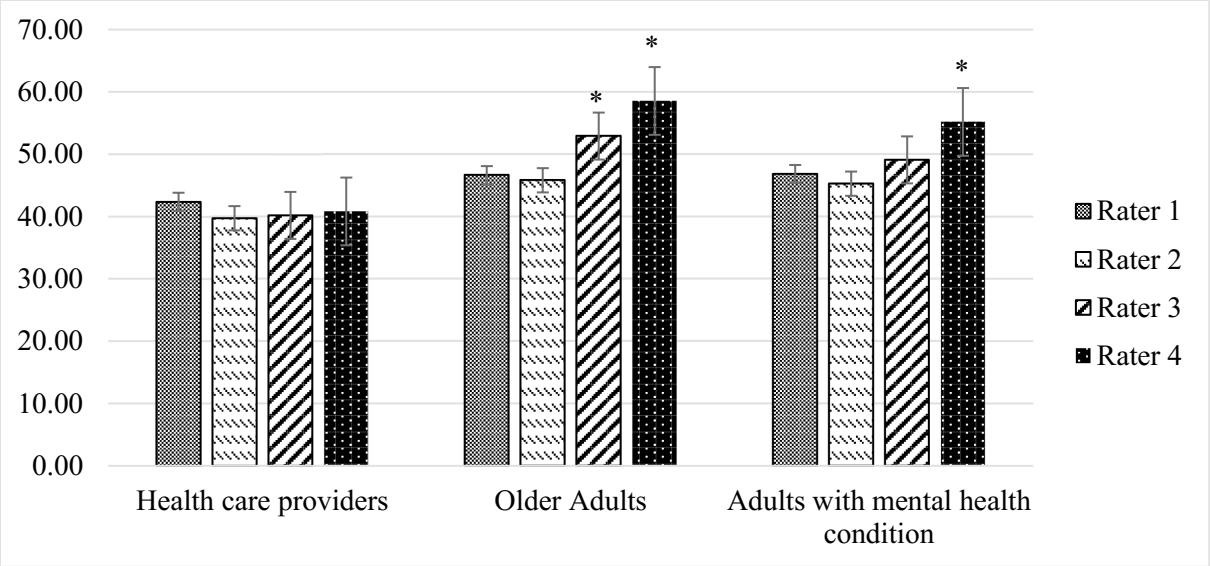

Figure 3. 8. Average total scores of ARIA by raters within each group.

* The mean score is significantly higher than the other raters within each group ($p<0.05$). Adjustment for multiple comparisons: Bonferroni.

On the other hand, while the total scores given by raters in each group positively correlated with their peers, the total scores given by rater 4 in the older adults group and the total scores given by rater 4 in the adults with mental health condition group negatively correlated with their peers (Table 3.11).

Table 3. 11. Correlation matrix of total scores of ARIA for all participants.

Group		Health care providers			Older Adults				Adults with mental health condition			
	Raters	2	3	4	1	2	3	4	1	2	3	4
Health care providers	1	0.898**	0.775**	0.862**	0.076	0.803**	0.679*	-0.676*	-0.516	-0.127	-0.468	0.614*
	2		0.942**	0.833**	0.007	0.843**	0.705*	-0.660*	-0.624*	-0.213	-0.584	0.587
	3			0.731*	0.001	0.722*	0.594	-0.514	-0.714*	-0.382	-0.718*	0.481
	4				0.201	0.845**	0.632*	-0.608*§	-0.263	0.129	-0.181	0.451
Older Adults	1					-0.043	0.643*	0.126	-0.004	-0.203	-0.094	-0.337
	2						0.566	-0.533§	-0.315	0.201	-0.213	0.412
	3							-0.449§	-0.432	-0.303	-0.471	0.235
	4								0.143	0.167	0.192	-0.881**§
Adults with mental health condition	1									0.506	0.930**	-0.232§
	2										0.765**	-0.126§
	3											-0.235§

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

§ Correlation with peers is negative.

Average ratings within items by raters ranged from 0.27 to 3.8 in the health care provider group (Table 3.12), 1.18 to 3.9 in the older adults group (Table 3.13), and 1.81 to 3.27 in the adults with mental health condition group (Table 3.14). The standard deviations around these averages were in the range of 0.405 and 2.00 that suggests ARIA was sensitive to variability in the different attributes of apps rated.

The negative correlations with peers suggest that rater 4 in the older adults group and rater 4 in the adults with mental health condition group may not have used ARIA the same way as their peers. One reason could be that they have misunderstood the instructions. Another reason could be that they have rated all apps high to please the researcher. As a result, the ratings by these two participants may negatively impact the reliability of the index and hence the investigator decided to remove the ratings of these two raters from the subsequent analyses in this study.

The investigator compared the average total scores between the iPhone users and Android users within each group using an un-paired t-test, to investigate any possible impact of the operating system used on the rating scores. No significant difference was found between iPhone users and Android users in the health care providers group. Interestingly, a significant difference existed between the mean total rating scores of iPhone users and Android users in the older adults group (mean difference = 9.5 points, 95% CIs [5.2, 13.7], $p<0.01$), and the adults with mental health condition group (mean difference = 6 points, 95% CIs [1.83, 10.35], $p< 0.01$). However, after removing the raters with off – the – chart ratings (i.e., rater number 4 in older adults group and rater number 4 in adults with mental health condition group) from the analysis, no significant differences were seen between the mean scores of iPhone users and Android users within any of the groups.

Table 3. 12. Ratings by items in health care providers group

	Health Care Providers Mean (SD)				
Items	Rater 1	Rater 2	Rater 3	Rater 4	All raters combined
A1	2.09 (2.02)	2.18 (2.09)	2.18 (2.09)	2.18 (1.33)	2.16 (1.84)
A2	2 (2)	2.45 (1.69)	2.18 (1.6)	2.18 (1.4)	2.2 (1.64)
A3	2.45 (1.69)	0.64 (1.12)	0.64 (1.12)	0.27 (0.9)	1 (1.48)
A4	0.36 (1.21)	0 (0)	0 (0)	0.36 (0.81)	0.18 (0.72)
A5	3.82 (0.6)	3.36 (1.21)	3.09 (1.58)	3.45 (1.21)	3.43 (1.19)
A6	3.45 (0.93)	3.55 (0.93)	3.55 (0.93)	3.18 (1.25)	3.43 (1)
B1	2.18 (2.09)	2.18 (2.09)	2.55 (2.02)	2 (1.95)	2.23 (1.98)
B2	0.91 (1.64)	0.73 (1.19)	0.45 (0.93)	2.64 (1.75)	1.18 (1.62)
B3	0.55 (1.29)	0.73 (1.27)	0.73 (1.27)	0.82 (1.08)	0.7 (1.19)
B4	3.27 (0.79)	3.36 (0.92)	3.36 (0.92)	2.82 (0.87)	3.2 (0.88)
B5	3.27 (0.65)	3.27 (0.65)	3.36 (0.5)	2.82 (0.6)	3.18 (0.62)
B6	3.45 (0.69)	3.27 (0.79)	3.45 (0.69)	3.45 (0.69)	3.41 (0.69)
B7	3.55 (0.52)	3.45 (0.93)	3.82 (0.4)	3.55 (0.52)	3.59 (0.62)
B8	1.82 (1.47)	1.55 (1.63)	2 (1.79)	1.55 (1.13)	1.73 (1.48)
B9	4 (0)	3.64 (0.5)	3.64 (0.5)	3.45 (0.52)	3.68 (0.47)
B10	1.73 (1.85)	1.82 (1.94)	1.73 (2)	2 (1.48)	1.82 (1.77)
B11	1.73 (1.85)	1.82 (1.94)	1.73 (2)	1.91 (1.58)	1.8 (1.79)
B12	1.73 (1.85)	1.73 (2)	1.73 (2)	2.18 (1.47)	1.84 (1.79)
Total	42.36 (8.94)	39.73 (8.81)	40.18 (8.69)	40.82 (5.88)	40.77 (7.95)
Stars	3.36 (1.36)	2.73 (1)	2.55 (1.21)	2.64 (0.92)	2.82 (1.15)

Table 3. 13. Ratings by items in older adults group

	Older Adults Mean (SD)				
Items	Rater 1	Rater 2	Rater 3	Rater 4	All raters combined
A1	3.45 (0.93)	1.27 (1.19)	3.64 (1.21)	3.27 (1.01)	2.91 (1.43)
A2	1.73 (1.19)	2.36 (1.36)	3 (1.1)	3.36 (0.81)	2.61 (1.26)
A3	1.27 (1.1)	1.18 (1.66)	2.27 (2)	3.18 (0.75)	1.98 (1.64)
A4	0.82 (0.4)	0.55 (1.04)	0.36 (0.67)	2.36 (1.12)	1.02 (1.15)
A5	3.64 (1.21)	3.73 (0.9)	3.55 (0.93)	3.36 (0.92)	3.57 (0.97)
A6	3.36 (1.21)	3.27 (1.1)	3.73 (0.9)	3.09 (1.04)	3.36 (1.06)
B1	2 (1.95)	2 (1.95)	2.18 (2.09)	2.09 (1.45)	2.07 (1.81)
B2	1.82 (1.4)	1.73 (1.62)	1.55 (1.86)	2.73 (1.27)	1.95 (1.57)
B3	0.73 (0.65)	1.91 (1.81)	1.45 (1.51)	3.09 (0.7)	1.8 (1.5)
B4	3.45 (0.52)	3.27 (0.65)	3.91 (0.3)	3.45 (1.21)	3.52 (0.76)
B5	3.09 (1.14)	3.55 (0.52)	3.73 (0.47)	3.55 (0.69)	3.48 (0.76)
B6	3.27 (0.47)	3.55 (0.52)	3.64 (0.81)	3.82 (0.4)	3.57 (0.59)
B7	3.27 (1.27)	4 (0)	4 (0)	3.91 (0.3)	3.8 (0.7)
B8	1.55 (1.21)	1.73 (1.19)	1.18 (1.4)	1.82 (1.47)	1.57 (1.3)
B9	3.27 (1.19)	3.64 (0.5)	4 (0)	3.91 (0.3)	3.7 (0.7)
B10	3.45 (1.29)	1.91 (1.51)	3.73 (0.9)	3.82 (0.6)	3.23 (1.34)
B11	3.18 (1.25)	3.18 (0.6)	3.64 (0.92)	3.91 (0.3)	3.48 (0.88)
B12	3.27 (1.19)	3 (0.63)	3.36 (0.81)	3.82 (0.4)	3.36 (0.84)
Total	46.64 (9.6)	45.82 (7.48)	52.91 (6.56)	58.55 (4.39)	50.98 (8.73)
Stars	3.09 (1.14)	3.27 (0.47)	3.82 (0.87)	4.45 (0.69)	3.66 (0.96)

Table 3. 14. Ratings by items in adults with mental health condition group

	Adults with mental health condition Mean (SD)				
Items	Rater 1	Rater 2	Rater 3	Rater 4	All raters combined
A1	2.91 (0.54)	2.36 (1.03)	3.09 (0.54)	2.45 (1.81)	2.7 (1.11)
A2	2.73 (1.01)	3 (0.45)	3.09 (0.54)	3.73 (0.65)	3.14 (0.77)
A3	1.27 (0.79)	2.18 (1.25)	2 (0.77)	1.82 (1.08)	1.82 (1.02)
A4	1.45 (0.69)	2 (1.1)	2 (0.77)	1.27 (1.01)	1.68 (0.93)
A5	3 (0.63)	3.18 (0.6)	3.09 (0.54)	3.64 (0.67)	3.23 (0.64)
A6	2.91 (0.94)	2.73 (1.01)	2.82 (0.87)	3.73 (0.9)	3.05 (0.99)
B1	2.27 (1.01)	2.45 (1.51)	2.55 (1.04)	2.09 (2.02)	2.34 (1.41)
B2	1.64 (0.5)	2.27 (0.9)	2.18 (0.6)	3.27 (1.1)	2.34 (0.99)
B3	2.64 (0.81)	2.45 (0.93)	2.73 (0.65)	3.82 (0.6)	2.91 (0.91)
B4	3.09 (0.54)	3.09 (0.3)	3.18 (0.4)	3.55 (0.93)	3.23 (0.6)
B5	3 (0.89)	3 (0)	2.91 (0.54)	3.73 (0.47)	3.16 (0.64)
B6	3.18 (0.75)	2.82 (0.75)	3.27 (0.65)	3.82 (0.4)	3.27 (0.73)
B7	3 (0.77)	2.82 (0.6)	3.09 (0.7)	3.73 (0.47)	3.16 (0.71)
B8	2.09 (1.04)	2 (1)	2.09 (0.7)	1.36 (1.36)	1.89 (1.06)
B9	3 (1)	2.45 (1.04)	2.91 (0.83)	3.36 (1.29)	2.93 (1.07)
B10	2.82 (0.75)	2.64 (0.5)	2.82 (0.6)	2.73 (1.85)	2.75 (1.04)
B11	2.91 (0.83)	2.18 (0.98)	2.64 (0.67)	3.55 (0.69)	2.82 (0.92)
B12	2.91 (0.83)	1.64 (1.29)	2.64 (0.81)	3.55 (1.21)	2.68 (1.23)
Total	46.82 (7.67)	45.27 (4.96)	49.09 (5.63)	55.18 (6.95)	49.09 (7.25)
Stars	3.27 (1.01)	2.73 (0.9)	3.27 (1.01)	4.18 (1.08)	3.36 (1.1)

3.3.2.3. The distribution of responses to the items

Figures 3.9 to 3.27 show the distribution of responses to the items of ARIA across all apps by each group when raters within each group are combined. Please note that raters who did not understand the task (i.e., rater four in the older adults group and rater four in the adults with mental health condition group) were excluded from this analysis.

In general, the shape of the distributions of the responses to the items were similar between three groups of raters (i.e., health care providers, older adults, and adults with mental health conditions). The distributions of responses to 1A (purpose), 2A (Trustworthiness), 5A (privacy), 6A (Affordability), 4B, 5B, 6B (Ease of use), 7B (functionality), and 9B (Appropriateness for target users) were negatively skewed. This finding indicates that raters in all groups scored these items higher for the apps that were included in this study.

In contrast, items such as 3A and 4A were positively skewed in all groups of raters, which means raters did not agree that the statements were true about most of the apps in the study. A closer examination of the app descriptions by the investigator indicated that many apps included in this study did not include the relevant information to help raters answer these items. For example, most of them did not have any statements about risks associated with using the app (topic of interest in item 3A) or none of them in fact clearly stated any conflicts of interests (topic of interest in item 4A) in the app descriptions or the developers' website.

The raters' responses to items such as 1B, 10B, 11B and 12B were distributed at the two ends of the scale (i.e., most responses were around 0 or 4). The shape of the distribution of responses to these items suggested that a dichotomous (Yes – No) scale may be a better alternative to rate these items.

Finally, the distribution of responses to the overall star ratings (Figure 3.27) indicated that, in general, the health care providers rated the apps lower than the older adults and adults with mental health condition group.

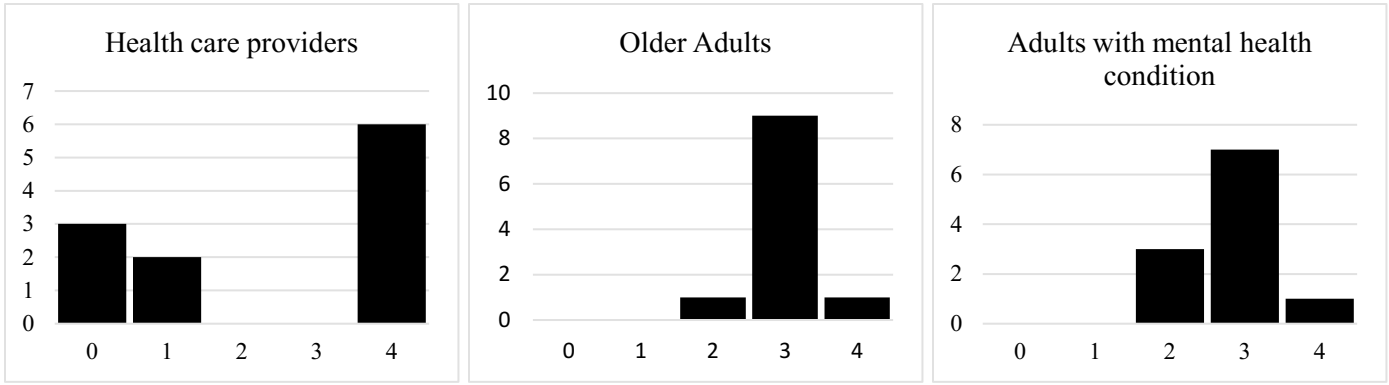

Figure 3. 9. Distribution of average rating scores to item 1A

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

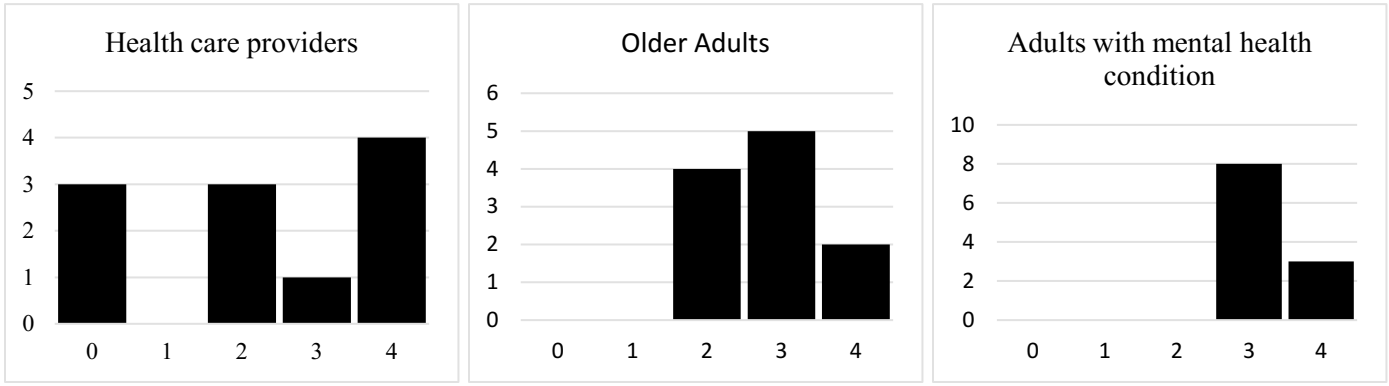

Figure 3. 10. Distribution of average rating scores to item 2A

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

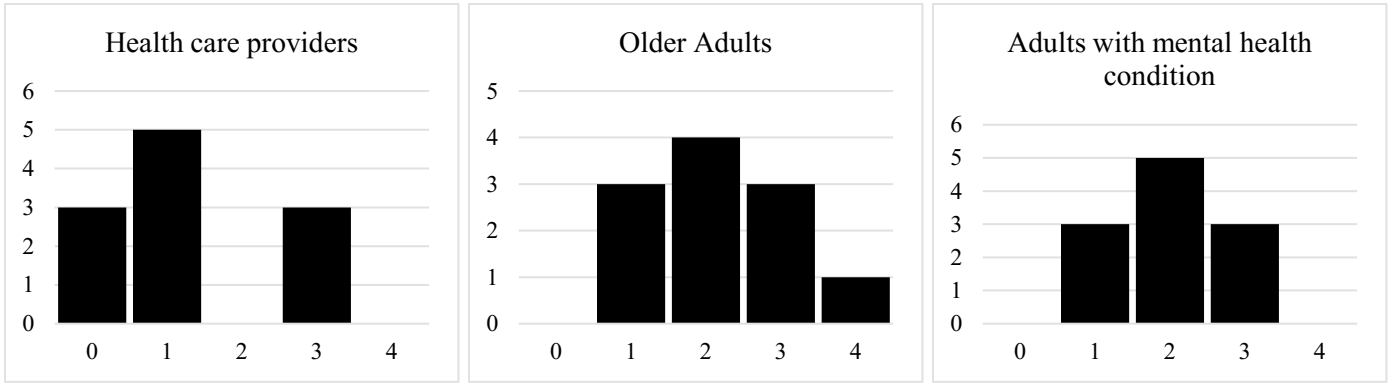

Figure 3. 11. Distribution of average rating scores to item 3A

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

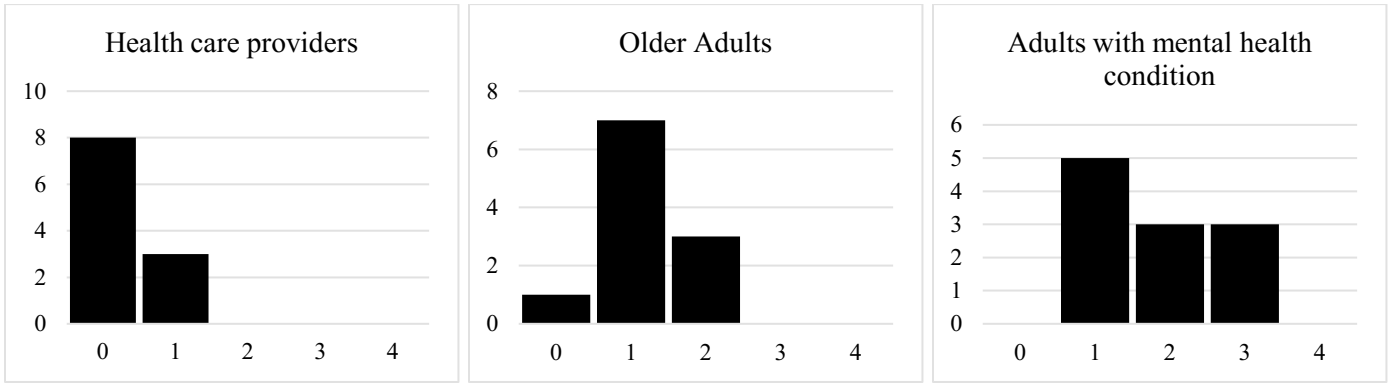

Figure 3. 12. Distribution of average rating scores to item 4A

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

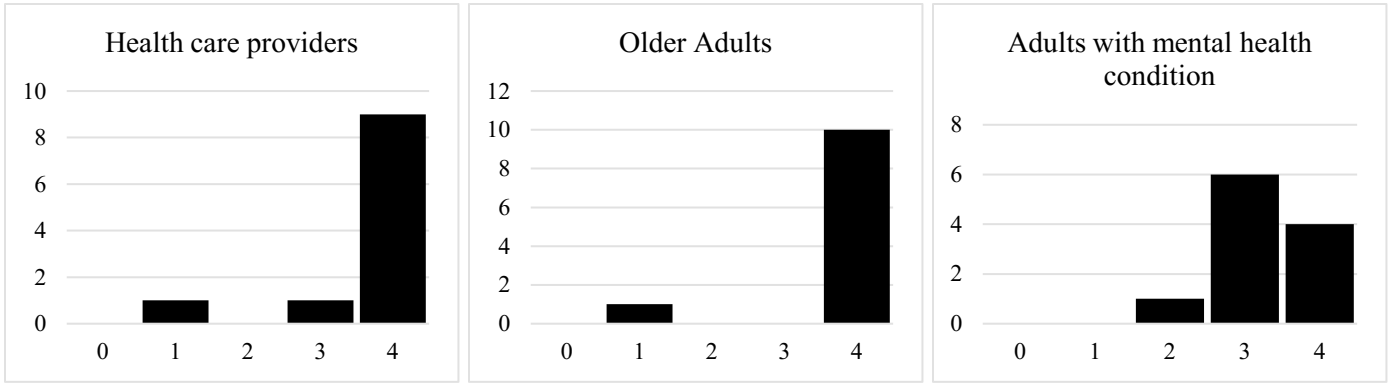

Figure 3. 13. Distribution of average rating scores to item 5A

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

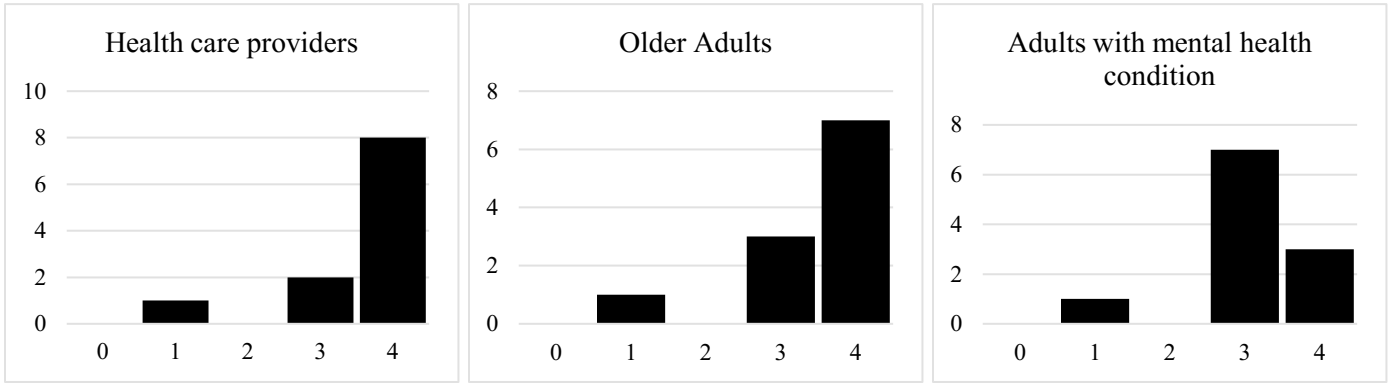

Figure 3. 14. Distribution of average rating scores to item 6A

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

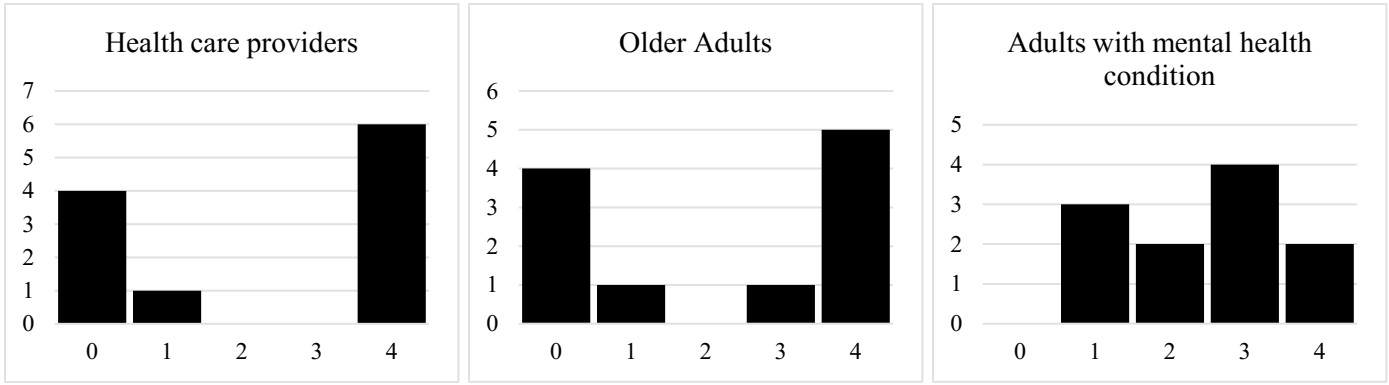

Figure 3. 15. Distribution of average rating scores to item 1B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

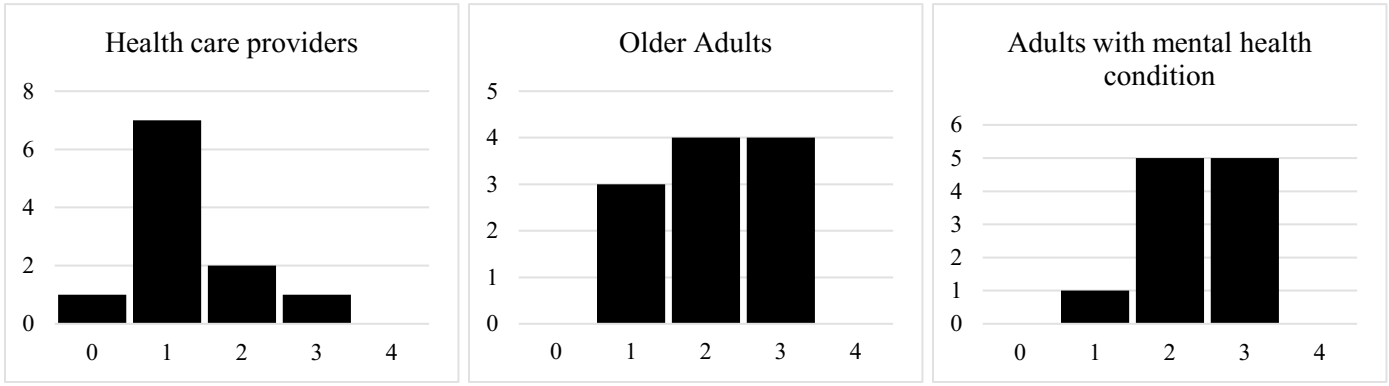

Figure 3. 16. Distribution of average rating scores to item 2B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

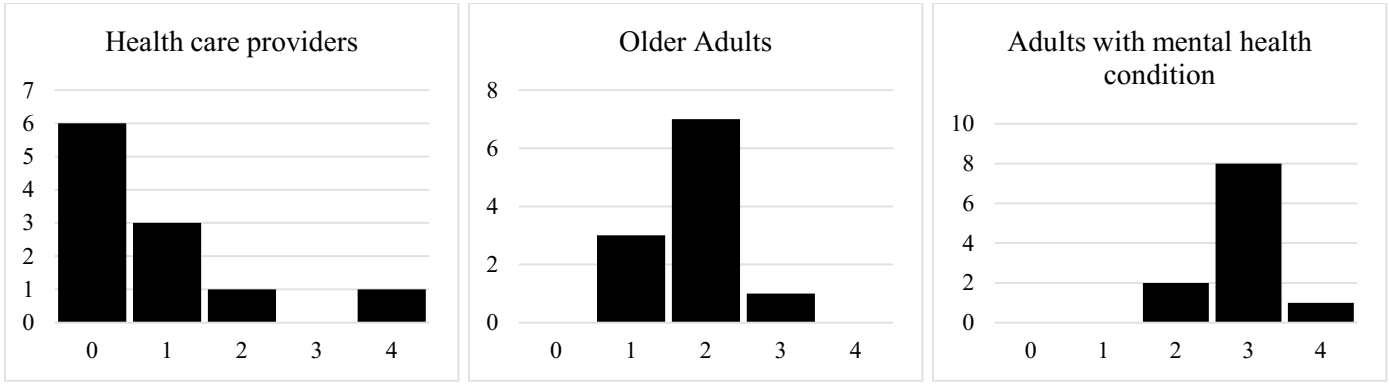

Figure 3. 17. Distribution of average rating scores to item 3B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

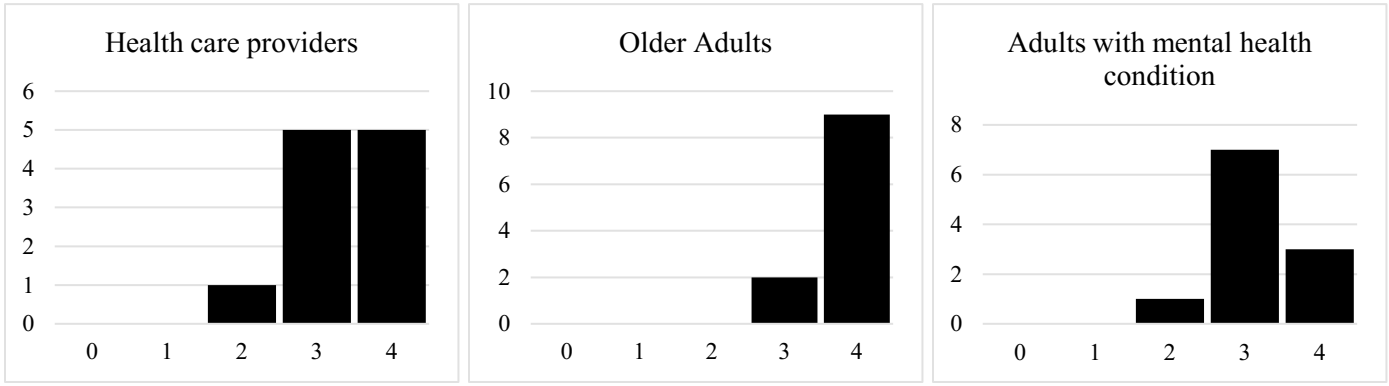

Figure 3. 18. Distribution of average rating scores to item 4B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

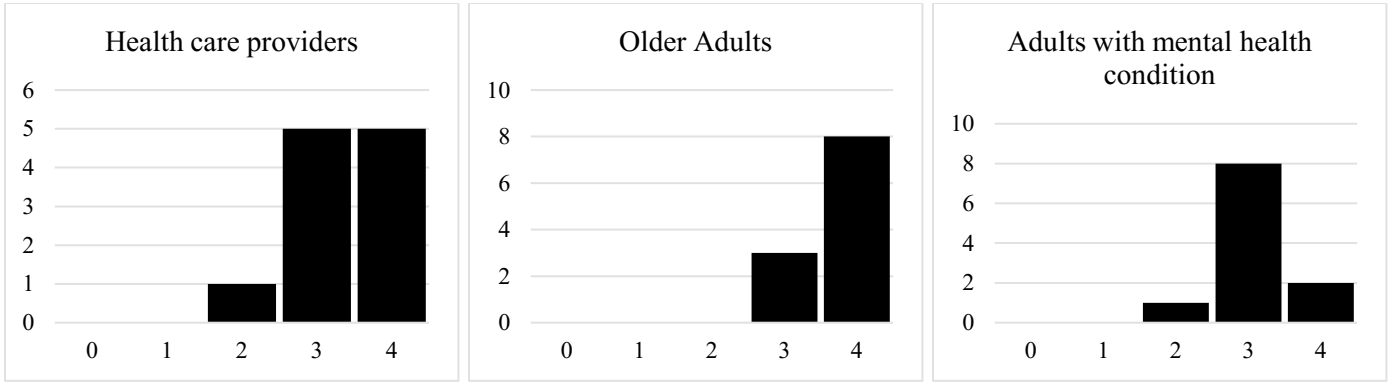

Figure 3. 19. Distribution of average rating scores to item 5B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

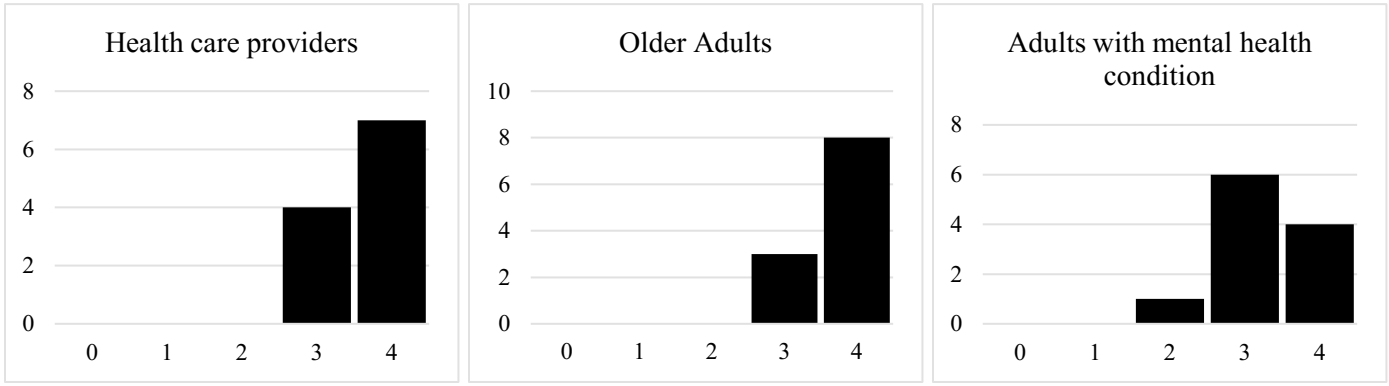

Figure 3. 20. Distribution of average rating scores to item 6B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

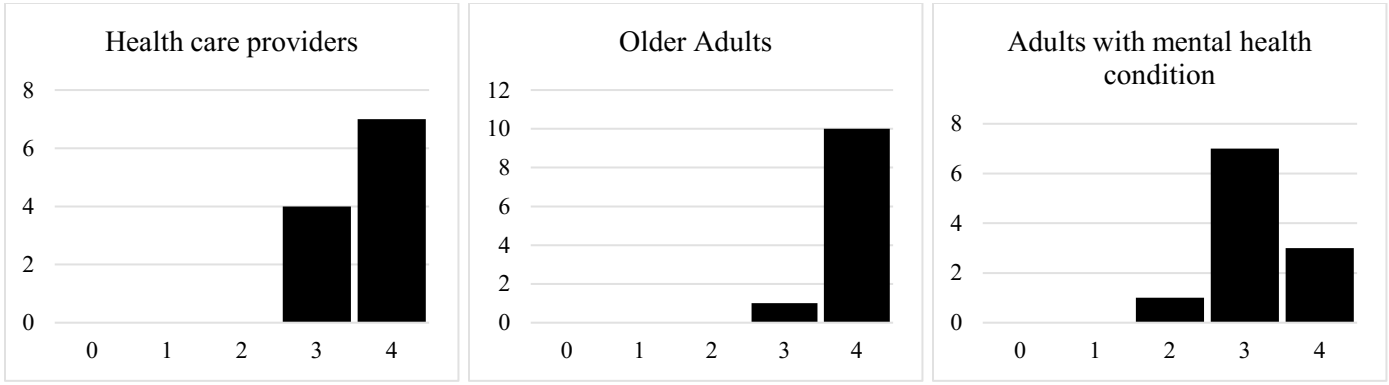

Figure 3. 21. Distribution of average rating scores to item 7B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

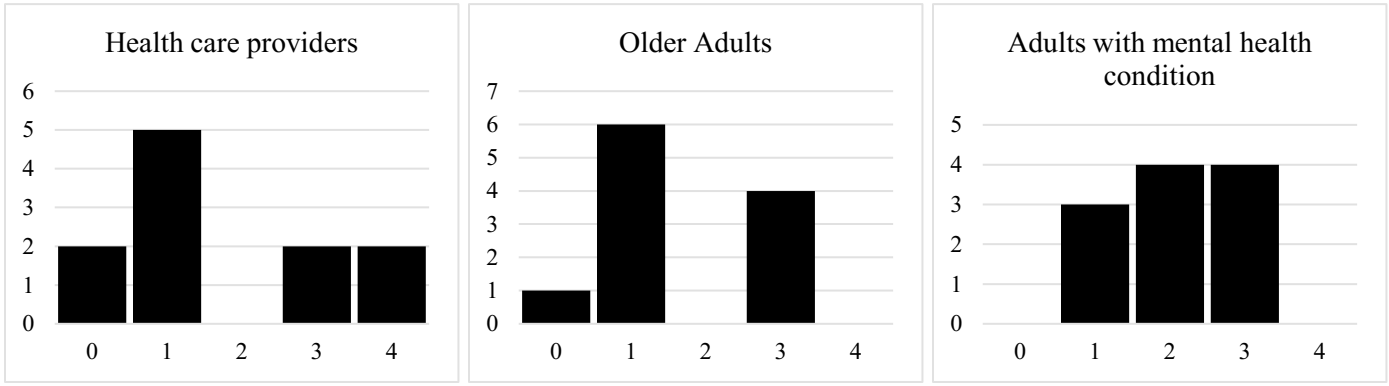

Figure 3. 22. Distribution of average rating scores to item 8B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

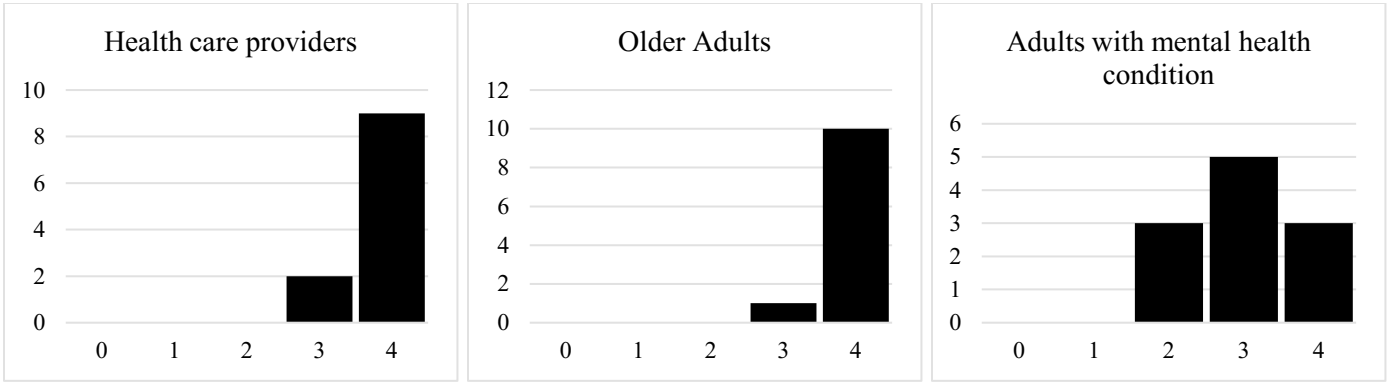

Figure 3. 23. Distribution of average rating scores to item 9B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

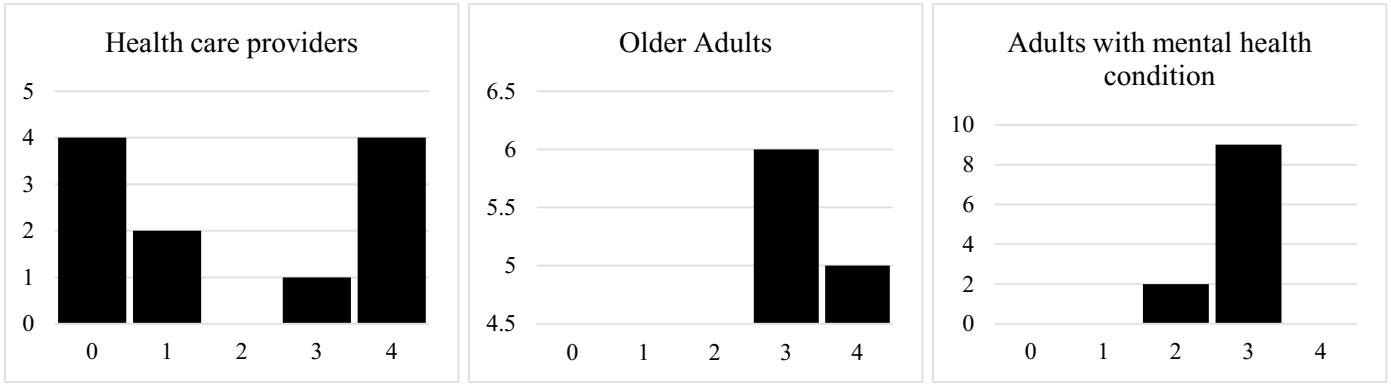

Figure 3. 24. Distribution of average rating scores to item 10B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

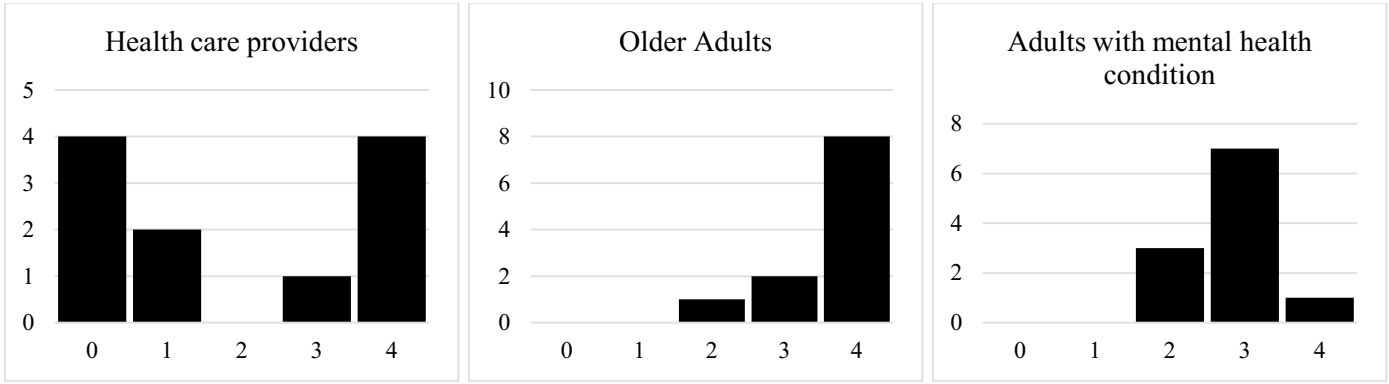

Figure 3. 25. Distribution of average rating scores to item 11B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

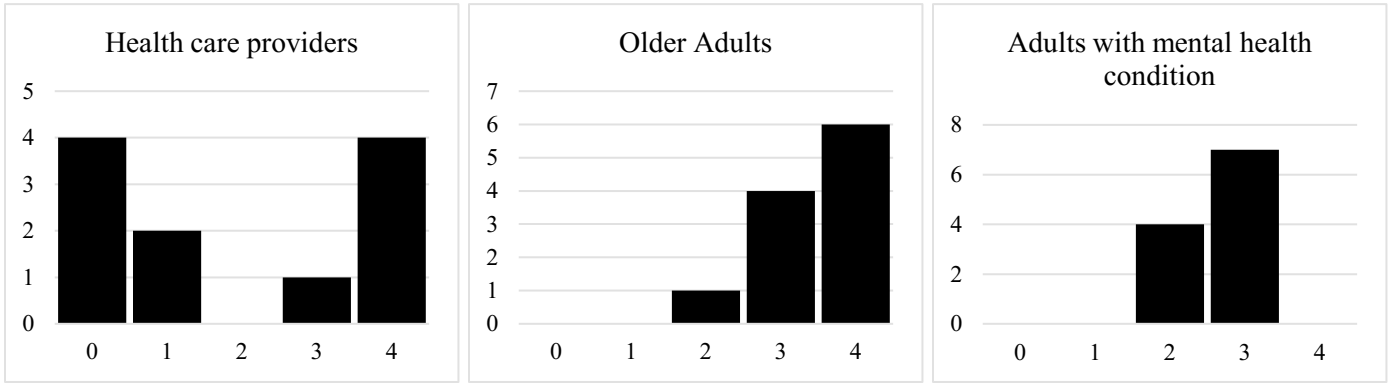

Figure 3. 26. Distribution of average rating scores to item 12B

Across all apps by each group; Raters within each group combined; ratings from 0 (Strongly Disagree) to 4 (Strongly Agree)

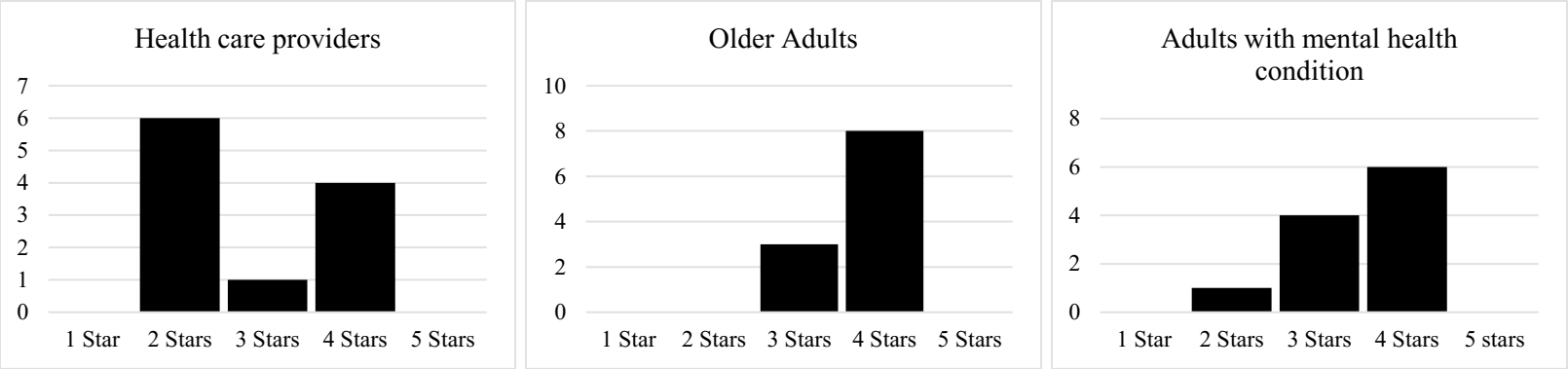

Figure 3. 27. Distribution of average star ratings across all apps by each group

(Raters within each group combined)

3.3.2.4. Inter-item correlations within each group

Tables 3.15 – 3.17 show the inter-item and item-total correlations by health care providers, older adults, and adults with mental health condition groups, respectively. For the older adults and adults with mental health condition, the correlations were calculated after removing the raters who did not understand the rating task (i.e., rater number four in older adults group and rater number four in adults with mental health condition group).

The investigator hypothesized that the quality of health apps was a composite variable and is defined by nine causal indicators (e.g., ease of use, security, affordability, etc.) that were not necessarily related to each other [101]. For example, if an app was easy to use, it did not mean that it was necessarily secure or affordable. As a result, the investigator did not expect to see a high correlation between the index item scores from different constructs (Tables 3.15 – 3.17).

Five constructs had more than one item. These were *Trustworthiness* (items 2A – 4A and Item 3B), *Security* (items 1B and 2B), *Ease of use* (items 4B – 6B), *Functionality* (items 7B and 8B), and *Usefulness and Satisfaction* (items 10B – 12B). Except for the *Ease of use* and *Usefulness and satisfaction*, there was a weak inter-item correlation between the items of the other constructs with multiple items.

The low inter-item correlation in items of *Security*, *Functionality*, and *Trustworthiness* may be due to the composite nature of these constructs. In other words, these constructs may have been consisted of items that are not necessarily correlated. For example, the items under security (1B: The app uses at least one security measure and 2B: The app asks for users' consent if it needs to access the phone's data) were completely independent from each other.

Tables 3.15 – 3.17 also present the correlation between each item and the total score as well as the correlation of the total score and the overall star ratings. Eight items in health care providers group, seven in older adults group, and 12 items in adults with mental health condition group correlated moderate to highly (0.5 – 0.8) with the index total score. The highest correlations were between items under *Usefulness and Satisfaction* and the total index score (Tables 3.17 – 3.19). The total index score and overall star ratings were also highly correlated in health care provider (0.843) and adults with mental health groups (0.759) and the older adults group (0.608).

Table 3. 15. Correlations between averages for item scores in Health care providers group

		Part A					Part B												
Items		2	3	4	5	6	1	2	3	4	5	6	7	8	9	10	11	12	Total	Stars
Part A	1	-.283	-.280	-.214	-.410	-.112	.544	.041	-.446	.606*	.606*	.462	.033	.300	.659*	.811**	.811**	.811**	.898**	.775**
	2		.366	.202	.091	.200	-.457	.085	.324	-.372	-.073	-.156	.031	-.526	-.544	.080	.080	.080	.067	.090
	3			.276	-.346	.497	-.415	-.203	-.236	-.129	-.086	.256	-.256	-.370	-.638	.000	.000	.000	-.255	.000
	4				-.191	.371	-.326	.112	-.534	.071	.285	.039	.039	-.682	-.241	-.238	-.238	-.034	-.260	-.109
	5					-.284	.355	-.180	.409	-.492	-.045	-.355	.177	.107	-.221	-.546	-.546	-.546	-.244	-.582
	6						-.146	.573	-.240	.423	.199	.305	-.076	-.135	-.285	-.079	-.079	-.079	-.123	-.150
Part B	1							.006	-.171	.287	.485	.134	-.034	.617*	.585	.126	.126	.126	.505	.134
	2								-.006	.269	.035	-.139	-.139	.246	.173	-.128	-.128	-.128	.019	-.172
	3									-.360	-.61	-.428	.066	.233	-.123	-.145	-.145	-.350	-.141	-.166
	4										.622*	.727*	.396	.064	.577	.454	.454	.359	.466	.394
	5											.727*	.529	-.234	.330	.322	.322	.322	.541	.254
	6												.607*	-.221	.134	.440	.440	.251	.361	.268
	7													-.442	.134	.000	.000	-.189	.060	-.134
	8														.433	.106	.106	.000	.244	.059
	9															.313	.313	.313	.525	.418
	10																1**	.91**	.871**	.938**
	11																	.91**	.871**	.938**
	12																		.813**	.938**
	Total																			.842**

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed)

Table 3. 16. Correlations between averages for item scores in older adults group

		Part A					Part B												
Items		2	3	4	5	6	1	2	3	4	5	6	7	8	9	10	11	12	Total	Stars
Part A	1	.323	.398	-.209	-.117	-.213	.141	.249	.141	.677*	.000	.339	-.175	.137	.151	.388	.200	.513	.401	.271
	2		.316	-.271	.106	.447	-.432	.013	.726*	.153	-.552	-.092	-.317	-.709	-.069	-.079	-.181	.133	.017	.092
	3			.300	.409	.142	-.123	.396	.458	.295	-.089	-.236	.038	.133	.198	.093	.381	.249	.746**	.295
	4				.280	-.090	.085	-.331	-.225	-.162	.162	-.162	.418	.243	-.181	.120	.085	.000	.332	.000
	5					-.235	.332	-.052	.221	-.289	.346	-.289	-.149	-.392	-.194	-.469	.221	-.442	.100	-.346
	6						-.424	.127	.224	.135	-.339	.000	.262	-.203	.076	-.038	.177	.430	.267	.339
Part B	1							.235	-.600	.000	.957**	.191	.000	.287	-.214	.318	.233	.150	.257	.000
	2								.052	.330	.240	-.090	-.077	.299	-.134	.310	.063	.261	.494	.180
	3									.191	-.638	-.128	-.082	-.569	.285	-.200	.067	-.094	.035	.000
	4										-.100	.267	.516	.162	.261	.339	.511	.670*	.492	.100
	5											.267	.043	.388	-.149	.271	.319	.128	.318	.100
	6												.043	.162	.261	.339	.319	.415	.145	.467
	7													.251	.241	.175	.577	.412	.374	-.043
	8														.326	.568	.276	.414	.490	.486
	9															.227	.642*	.392	.291	.559
	10																.094	.731*	.556	.677*
	11																	.511	.630*	.319
	12																		.725*	0.702*
	Total																			0.608*

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed)

Table 3. 17. Correlations between averages for item scores in adults with mental health condition group

		Part A					Part B												
Items		2	3	4	5	6	1	2	3	4	5	6	7	8	9	10	11	12	Total	Stars
Part A	1	.330	-.216	-.248	-.041	.289	.099	-.381	.335	.381	.374	.748**	.773**	-.263	.303	.313	.302	.416	.499	.610*
	2		-.308	.000	.330	.773**	.078	-.075	.628*	.856**	.707*	.543	.542	-.006	.303	.347	.622*	.589	.687*	.610*
	3			.641*	-.524	-.381	-.085	.453	-.433	-.396	-.277	-.650*	-.408	-.021	-.432	.198	-.240	-.024	-.149	-.157
	4				.224	.064	-.141	.639*	-.218	.000	.000	-.218	-.022	.000	-.384	.456	-.215	.000	.335	.000
	5					.289	.214	-.075	.400	.381	.374	.543	.484	-.035	.303	.381	.141	.106	.593	.270
	6						-.380	.321	.511	.736**	.449	.539	.427	.207	.107	.123	.452	.261	.580	.440
Part B	1							-.476	.267	-.039	-.078	.077	.172	-.204	-.156	.149	-.136	.070	-.038	.173
	2								-.102	-.175	-.265	-.406	-.301	.200	-.533	-.175	-.330	-.452	-.038	-.279
	3									.543	.218	.483	.496	.242	.237	.269	.592	.365	.541	.474
	4										.523	.585	.659*	.288	.390	.524	.577	.534	.679*	.433
	5											.628*	.552	-.542	.610*	.265	.500	.598	.681*	.607*
	6												.825**	-.242	.474	.347	.456	.478	.722*	.719*
	7													-.206	.470	.483	.244	.330	.732*	.473
	8														-.076	.247	.217	-.083	-.056	-.210
	9															.310	.525	.378	.454	.276
	10																.471	.617*	.755**	.526
	11																	.859**	.610*	.701*
	12																		.644*	.780**
	Total																			.759**

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

3.3.2.5. Generalizability analysis of the items of ARIA

Tables 3.18 – 3.20 present the variance components and the proportions of variance (Proport.) from the two facet (A x I x R) G-study analysis of the items (I) and raters (R) in health care providers, older adults, and adults with mental health condition groups respectively. The proportions of variance associated with raters (R) is negligible is all groups (0.00 in health care providers, 0.04 in older adults, and 0.04 in adults with mental health condition), which indicates a high level of consistency among raters. Moreover, a high proportion of variance associated with items (0.17 in older adults, 0.30 in adults with mental health condition, and 0.34 in health care providers) indicates that items of ARIA measure criteria that are not related to each other.

Table 3. 18. Variance decomposition of items (health care providers)

	df	**SS**	**MS**	**Variance**	**Proport.**	**G**
A	10	128.513	12.851	0.099	0.033	0.552
I	17	878.864	51.698	1.023	0.343	
R	3	2.429	0.81	0	0	
A x I	170	945.942	5.564	1.271	0.426	
A x R	30	20.154	0.672	0.011	0.004	
I x R	51	81.753	1.603	0.102	0.034	
A x I x R , e	510	244.664	0.48	0.48	0.161	

df: degree of freedom; SS: sum of square, MS: mean square, G: Measure of reliability
A: Quality of apps (Object of measurement); I: Index items (Facet 1);
R: Raters (Facet 2)

The proportion of variance (Proport.) associated with app quality (A) was notably low in the older adults group (0.00) and adults with mental health group (0.011) in comparison to the health care providers group (0.033). This is also reflected by the low generalizability coefficient (G) in older adults (0.002) and adults with mental health condition (0.248) compared to the health care providers group (0.552).

Table 3. 19. Variance decomposition of items (older adults)

	df	SS	MS	Variance	Proport.	G
A	10	40.402	4.04	0	0	0.002
I	17	567.052	33.356	0.66	0.307	
R	3	65.064	21.688	0.086	0.04	
A x I	170	376.962	2.217	0.369	0.172	
A x R	30	76.588	2.553	0.101	0.047	
I x R	51	145.095	2.845	0.191	0.089	
A x I x R , e	510	377.503	0.74	0.74	0.345	

df: degree of freedom; SS: sum of square, MS: mean square, G: Measure of reliability
A: Quality of apps (Object of measurement); I: Index items (Facet 1);
R: Raters (Facet 2)

Table 3. 20. Variance decomposition of items (adults with mental health condition)

	df	SS	MS	Variance	Proport	G
A	10	37.924	3.792	0.013	0.011	0.248
I	17	195.818	11.519	0.201	0.172	
R	3	34.737	11.579	0.044	0.038	
A x I	170	263.848	1.552	0.272	0.233	
A x R	30	52.874	1.762	0.072	0.062	
I x R	51	81.444	1.597	0.103	0.088	
A x I x R, e	510	236.444	0.464	0.464	0.397	

df: degree of freedom; SS: sum of square, MS: mean square, G: Measure of reliability
A: Quality of apps (Object of measurement); I: Index items (Facet 1);
R: Raters (Facet 2)

The variance proportion associated with the error interaction between the quality of apps, Items, and Raters (A x I x R, e) was relatively high especially in older adults (0.345) and adults with mental health condition (0.394). This finding suggests that there might be potential rater idiosyncrasies and other factors such as unidentified facets and confounding effects that influence ratings of the quality of m-health apps, especially by older adults and adults with mental health conditions [153]. This finding is in accord with the observed low G coefficient and low proportion of variance associated with app quality (A) in older adults and adults with mental health condition groups.

The variance proportion associated with error interaction was much lower in the health care providers group (0.161), which is in accord with the higher variance proportion associated with apps quality (A) and the higher G coefficient in the health care provider group. A possible explanation can be that the health care providers in this study rated apps according to a very detailed patient profile which reduced the amount of unaccounted variabilities and idiosyncrasies (i.e., different needs and expectations) that existed among participants in older adults group and adults with mental health condition group.

The generalizability analysis of the app quality based on the item scores confirmed two points. First, the high variance proportion associated with the items provides supporting evidence that items of ARIA measure different attributes of apps. Second, the users' responses to these items can be influenced by their personal needs, experience, and expectations, which was reflected by high interaction (A x I x R, e) and low reliability coefficient (G). This indicates that items of ARIA enable users to rate quality of apps based on their own needs and expectations. As it was observed in the health care providers group, when the raters are very similar in terms of their needs and expectations, the interaction error term will decrease and the reliability coefficient (G) will increase.

3.3.2.6. Generalizability analysis of total score of ARIA

A single facet G-study was conducted on apps and raters (App x R), using the total scores of ARIA, for each rater group separately to investigate the reliability of the composite score that combined all individual items in the index.

3.3.2.6.1. Results from the health care providers group

Variance components and the proportions of the variance of the G-study in Health care provider group is presented in table 3.21. The proportion of variance associated with apps (APP),

raters (R), and error interaction (APP x R, e) are 0.816, 0.003, and 0.18, respectively. The variance proportion associated with apps indicates a large proportion of the observed variability (81.6%) is associated with the differences between app qualities. Moreover, the variance associated with differences between raters is negligible (0.003) that means rating behavior did not substantially vary between raters in health care providers group. Finally, a high G-Coefficient of 0.948 indicates that the total score of ARIA can be used reliably to measure the quality of apps by health care providers.

Table 3. 21. Variance decomposition of total score (health care providers)

Source	df	SS	MS	Variance	Proportion of Variance explained	G
APP	10	2313.227	231.323	54.808	0.816	0.948
R	3	43.727	14.576	0.226	0.003	
APP x R, e	30	362.773	12.092	12.092	0.18	

APP: quality of apps (object of measurement); R: Raters (Facet 1), e: error

3.3.2.6.2. Results from the older adults group

The G-study of index total scores from the older adults group is presented in Table 3.22. Analysis of data from all participants (condition 1) showed a large proportion of variance associated with error interaction (0.547) and a very low G-coefficient (0.368). As indicated earlier, the investigator suspected that one of the raters (rater 4) did not completely understand the task and as a result produced inconsistent ratings in comparison to his peers. Removing this participant from the analysis resulted in a significant increase in the variance proportion associated with the index scores (+0.21) and a significant increase in the G-coefficient (+0.36).

The investigator noticed that one of the participants (rater 1) in the older adults group had rated one of the apps (*Headspace*) very low in comparison to his peers (19 versus 50 average score of the other three). To remove the confounding effect of this error, the investigator excluded the scores for the *Headspace app* from the older adults group and ran the G-study once more (condition 3). The results indicated a significantly higher G-coefficient (0.829), as well as high variance proportions associated with total sores (0.419), and a much smaller variance proportion associated with error interaction (0.28).

Table 3. 22. Variance decomposition of total score (older Adults)

Condition	Source	df	SS	MS	Variance	Proportion of Variance explained	G
1	APP	10	727.227	72.723	6.692	0.08	0.368
	R	3	1171.16	390.386	31.312	0.373	
	APP x R, e	30	1378.59	45.953	45.953	0.547	
2	APP	10	1072.85	107.285	21.758	0.29	0.608
	R	2	331.091	165.545	11.23	0.15	
	APP x R, e	20	840.242	42.012	42.012	0.56	
3	APP	9	712.167	79.13	21.859	0.419	0.829*
	R	2	361.4	180.7	16.715	0.321	
	APP x R, e	18	243.933	13.552	13.552	0.26	

APP: Quality of apps (object of measurement); R: Raters (Facet 1), e: error
Condition 1: 4 raters and 11 apps
Condition 2: 3 raters and 11 apps (excluding rater number 4)
Condition 3: 3 raters and 10 apps (excluding rater number 4 and app number 11)
*Highest G coefficient

The proportion of variance associated with raters was relatively high among older adults in all three conditions (0.373, 0.15, and 0.321) which indicates the personal differences between the participants in this group and its impact on the ratings.

3.3.2.6.3. Results from the adults with mental health condition group

Like the older adults group, the investigator ran the analysis for the adults with mental health condition group once with all raters included and once after removing rater 4. By removing rater 4, the variance proportion associated with index total score increased from 0.16 to 0.656. Besides the variance proportion associated with interaction error term decreased from 0.55 to 0.279, and the variance proportion associated with rater effect decreased from 0.282 to 0.065. Finally, the G-coefficient increased from 0.535 to 0.876 (Table 3.23).

Table 3. 23. Variance decomposition of total score (adults with mental health condition)

Condition	Source	df	SS	MS	Variance	Proportion of Variance explained	G
1	APP	10	682.636	68.264	9.135	0.16	0.535
	R	3	625.273	208.424	16.064	0.282	
	APP x R x e	30	951.727	31.724	31.724	0.557	
2	APP	10	921.879	92.188	26.915	0.656	0.876*
	R	2	81.152	40.576	2.648	0.065	
	APP x R x e	20	228.848	11.442	11.442	0.279	

APP: Quality of apps (object of measurement); R: Raters (Facet 1), e: error
Condition 1: 4 raters and 11 apps
Condition 2: 3 raters and 11 apps (excluding rater number 4)
*Highest G-coefficient

3.3.2.7. Generalizability study of the overall star ratings

A single facet G-study was conducted on the quality of apps and the raters (APP x R) using the overall star ratings for each group. In the health care providers group the results of the G-study using the overall star ratings (Table 3.24) were consistent with the results of the G-study using the total index score: high variance proportion for the object of measurement or the quality

of apps (0.576), low variance proportion for raters (0.066) and interaction (0.358), and a high G-coefficient (0.865).

Table 3. 24. Variance decomposition of star ratings (health care providers)

Source	df	SS	MS	Variance	Proportion of Variance explained	G
APP	10	37.045	3.705	0.802	0.576	0.865
R	3	4.545	1.515	0.092	0.066	
APP x R, e	30	14.955	0.498	0.498	0.358	

APP: The quality of apps (object of measurement); R: Raters (Facet 1), e: error

However, this was not true for the older adults group. Unlike the health care providers group, the results of G-study of overall star rating for older adults showed a G-coefficient of 0 and a very large variance proportion associated with rater x total score interaction (0.706), which indicates that the older adult participants in this study did not reliably use the star rating system to rate the quality of apps (Table 3.25).

Table 3. 25. Variance decomposition of star ratings (older adults)

Source	df	SS	MS	Variance	Proportion of Variance explained	G
APP	10	5.136	0.514	0	0	0
R	3	12.432	4.144	0.309	0.294	
APP x R, e	30	22.318	0.744	0.744	0.706	

APP: The quality of apps (object of measurement); R: Raters (Facet 1), e: error

The results from adults with mental health condition followed the same pattern as the health care providers group. However, a larger variance proportion associated with the raters (0.226) and APP x Rater interaction error was observed in this group (0.59). One possible explanation is that unlike the health care provider group who rated the apps based on the same mock patient profile, participants in the adults with mental health condition group rated the apps based on their own needs. Therefore, the variance proportion associated with the interaction of rater and scores reflects the personal differences between patients when evaluating overall app quality using star ratings (Table 3.26). This difference has also been reflected in the lower G-coefficient in adults with mental health condition group in comparison to the health care providers group.

Table 3. 26. Variance decomposition of star ratings (adults with mental health condition)

Source	df	SS	MS	Variance	Proportion of Variance explained	G
S	10	17.182	1.718	0.238	0.183	0.554
R	3	12	4	0.294	0.226	
S x R, e	30	23	0.767	0.767	0.59	

S: star rating (object of measurement); R: Raters (Facet 1), e: error

3.3.3. Validity testing

To investigate the criterion validity of ARIA, the average total scores of ARIA given by the participants in each group were correlated with the average total scores from the users' version of Mobile Application Rating Scale (U-MARS) for the same apps given by the same participants. Although U-MARS is not the gold standard app quality scale, it is currently the only one with a users' version available.

The results are presented in Figures 3.28 – 3.30. The Spearman's rho correlation test indicated a low to moderate positive correlation between total scores of ARIA and U-MARS repeated over all three groups of participants. The results indicate that ARIA and U-MARS measure similar but different attributes of an underlying concept that could be associated with the quality of mobile health apps.

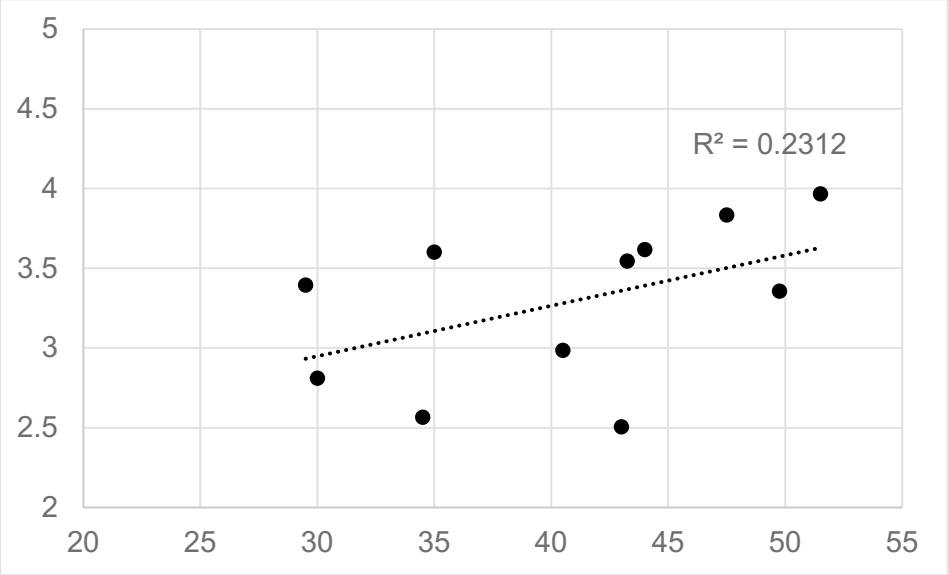

Figure 3. 28. Correlation between ARIA and U-MARS (health care providers)
(Spearman's rho for total scores = 0.526)

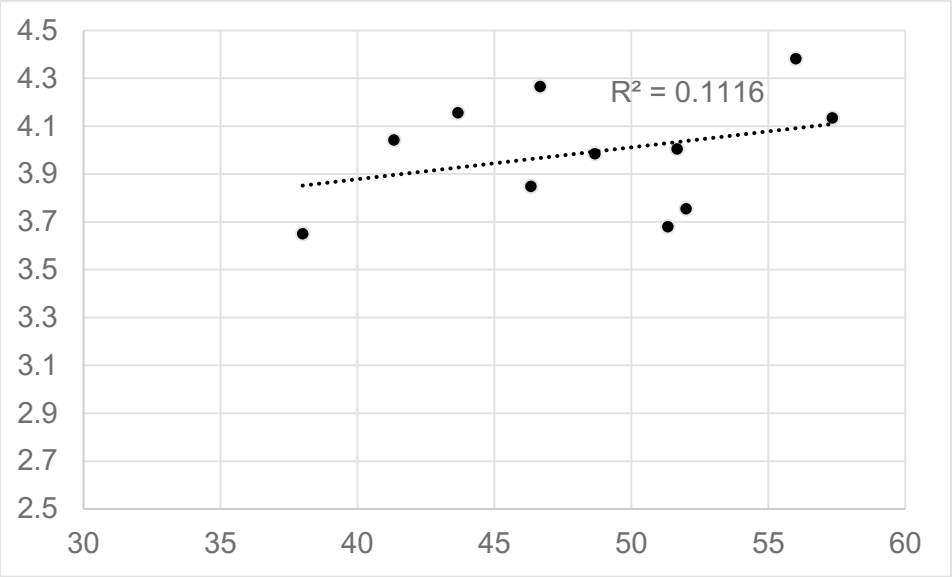

Figure 3. 29. Correlation between ARIA and U-MARS (older adults)

(Spearman's rho for total scores = 0.128)

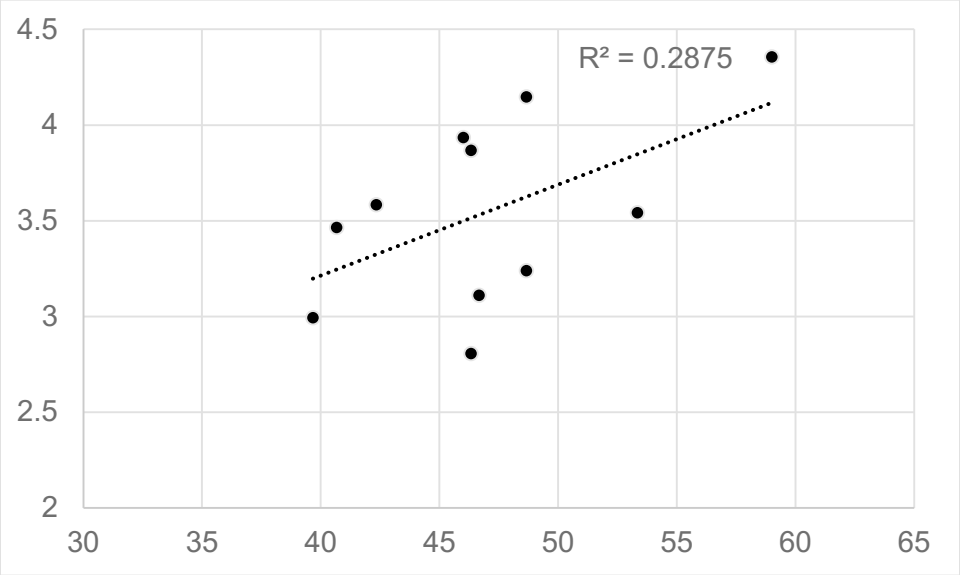

Figure 3. 30. Correlation between ARIA and U-MARS (adults with mental health condition)

(Spearman's rho for total scores = 0.38)

3.3.4. Raters' comments.

At the end of the reliability study, the investigator asked participants who used both ARIA and U-MARS share their feedback and comments about using these assessment tools to rate the sample of apps in this study.

All participants said that they found ARIA easier to use than U-MARS. Participants mentioned that they found ARIA shorter, easier, and more comprehensive. Moreover, they said that ARIA was more user-centered and enabled them to assess the app based on their needs and objectives.

Older adult 1: *"I think it has the right questions, and I think it covers all the essential points, and it is easy to use."*

Older adult 2: *For example, for me I would have scored the medication reminder app fairly highly [with the new index] because the app was relevant to me but other apps would not, but when I get to the U-MARS scoring I might not have given it [medication reminder] a score higher than the other apps because I was rating it against whether I thought that app would be useful for somebody else, with all those functions, so it did not lose or gain any points for relevance to me.*

Older adult *2: The value of the AGE-WELL one [ARIA] was that it was short, and it just required you to add up things. The U-MARS was also relatively easy, but I needed to pull out the calculator from time to time, and sometimes it was four questions, sometimes three, and five and so the chances of error increased.*

Health care provider 1: *"Definitely yours [ARIA] is easier since it does not require me to calculate the averages [unlike in U-MARS], and it is easier to understand because you have just 5-point Likert scale and at the end, I just have to add the scores and write it down".*

Adult with mental health condition 2: *"The U-MARS was also good, but sometimes my math was probably off, I always had to go back and check if it was right."*

Participants in the older adults group and health care providers group said that it was difficult to answer item 3A (availability of a statement on risks) and 4A (if the app states any conflicts of interest), based on the information available on the app stores or within the apps.

Health care provider 4: *"The other one was number 3 the risk association, I had a hard time finding if there was any risk associated".*

Older adult 2*: "With the AGE-WELL form [ARIA], often apps missed any indication of potential risks associated with it, and that was a question that caused me rating them very low. The conflict of interest was also seldom listed in either, but even, later on, I had to dig to try and find any indication".*

Health care provider 1: *"The question number 4 in part A is complicated. I would suggest change or remove it."*

The negatively skewed distribution of the scores to item 3A and 4A (see 3.3.2.4) confirms this comment by the participants and supports that many apps fail to include information about the risk associated with their use or any conflict of interest. These findings may suggest that such items be removed or reworded.

Participants believed that some items might not apply to certain apps and hence, the not applicable option may help in those situations. However, they also agreed that adding a not applicable option can make the total score calculation cumbersome.

Older adult 2: *"Security was an interesting question especially B2 question it is not always an issue if the app is asking your personal information. I do not mind not having a password if it is an app that is helping me meditate, because it is not gathering any information,*

I just put it down and cool my jets. It might be a little bit difficult for me to evaluate if it is asking to rate it between 0 and 4, but how do I rate it if security is not an issue. If it did not ask me for security and I gave it 0, versus when it did not need any security, so I gave it a 4. It is one of those ones that having an NA would be valuable and somehow you will need to exclude that number which implied to me that it might be better for the overall scoring to do a percentage score or something so that I can adjust the total number and exclude some questions because they were not relevant to that particular app".

Finally, raters found ARIA educational and that it brought to their attention aspects of the apps that they did not consider previously.

Health care provider 3: *"I liked your cues because being new at it not knowing where to find the information, not knowing what to look for, I really enjoyed having those cues to guide me to this is where you need to look for, so I thought those were really helpful."*

Overall, participants' feedback and comments regarding their experience using ARIA to rate the quality of the apps were positive and supported our findings from the quantitative analysis.

CHAPTER 4: DISCUSSION AND CONCLUSION

4.1. Project Overview

The objective of this study was to develop the Alberta Rating Index for Apps (ARIA), a rating index for users of m-health apps such as the general public, patients, older adults, family caregivers, and health care providers to make informed decisions when they choose m-health apps.

To achieve this objective, a multi-strategy study was conducted in three phases to answer the following research questions:

1) What are the criteria to rate the quality of mobile health apps?
2) How can users of m-health apps measure the extent to which m-health apps meet the quality criteria mentioned in question one?
3) Can health care providers, older adults, and adult with a chronic health condition use ARIA, a criteria-based index, to reliably rate the quality of m-health apps?
4) What is the criterion validity of total quality scores of ARIA against U-MARS?

In phase one, the criteria that were important to measure the quality of m-health apps were identified. First, the investigator extracted the quality criteria from three models of technology acceptance and use and four frameworks of mobile app quality evaluation. Next, different stakeholders, including older adults, adults with mental health conditions, clinicians, and app developers, participated in 6 focus groups, to validate the identified quality criteria or suggest new ones that were missing. At the end of this process, the investigator developed a user based framework for evaluating m-health apps based on the criteria that were approved by the stakeholders.

In the second part of phase one, an item pool was generated to measure the quality criteria identified in the first part. The investigator used a systematic review of the literature to locate scales and assessment tools previously developed to rate the quality and usability of information technologies and mobile applications. Next, scale items were extracted from the scales identified. A panel of three experts in the field of health technology evaluation mapped the items to the quality criteria identified in the first part. Unnecessary items were removed and similar ones were merged. The result was an item pool with 74 items distributed unevenly under 10 quality criteria.

In phase two, the investigator validated the content of the item pool. He used an online survey to recruit 18 stakeholders and ask them to review the item pool and rate the relevance of each item to the given criteria. Participants in this study included app evaluators, clinicians, older adults, and adults with mental health conditions. Participants were recruited from four different provinces of Canada to increase the diversity of opinions involved in the content validation process. A content validity index (CVI) was calculated based on the responses from the participants. The investigator kept items with CVI $\geq$ 80 (endorsed as valid by at least 80% of the survey participants) and deleted the rest of the items from the item pool. In the next step, one-third of the participants in the online survey (n=6) joined in a focus group to shortlist the items identified as relevant in the online survey and create the first draft of the index. The primary objective of this activity was to create an index with the right length (i.e., number of items) to be usable by the general public while it covered all the quality criteria identified in phase one. The first draft of the index included 18 items, nine criteria, and two sections and was named Alberta Rating Index for Apps (ARIA).

In phase three, first the investigator pilot-tested the pilot draft of ARIA with nine users of m-health apps (two older adults, and seven adults with mental health conditions). He revised pilot draft of ARIA based on the comments and feedback from participants in the pilot study. In the second part of phase three, final version of ARIA was tested by 12 potential users of m-health apps on 11 different health apps mostly related to mental health conditions. The apps were selected form the directory of mental health apps of Alberta health services (AHS) addiction and mental health research partnership program [28]. Participants were requested to download each app on their smartphone, use the app for as long as they need to become familiar with its features, and then rate the quality of each app twice: once with our index (ARIA) and once with the users' version of Mobile Application Rating Scale (U-MARS). The investigator used the ratings to investigate the inter-rater reliability of the index in a Generalizability study (G-Study). The investigator also correlated total scores of ARIA with total scores of U-MARS to study the criterion related validity of ARIA.

4.2. Quality Criteria for M-Health Apps

The first research question was to identify the criteria for evaluating the quality of apps. To answer the first research question the investigator reviewed the current scales used in the literature for evaluation of m-health apps as well as all theories and models for the usability of information technology and evaluation of the quality of mobile apps.

The review indicated that there are at least 48 different scales used in the literature to evaluate mobile health apps, from which 23 scales were used for usability evaluation of apps. Usability refers to a multi-dimensional property of a computer interface that is associated with attributes such as ease of use, learnability, efficiency, memorability, error rates, and user satisfaction [33]. Such criteria are usually of interest to app developers and human-computer

interface researchers. Most of these criteria are common between mobile apps and other information technologies, such as computer programs and websites. Several theoretical models have been developed to explain and predict consumers' behavior in terms of acceptance and use of information technologies. Examples are the Technology Acceptance Model (TAM) [88], Rogers' Innovation Diffusion Theory [89], and the revised Unified Theory of Acceptance and Use of Technology (UTAUT 2) [92].

While usability is a necessary condition, it is not sufficient for quality of m-health apps. Other factors such as evidence of scientific accuracy, the effectiveness of the interventions, and safety have been reported in the literature as additional quality criteria for health information technologies including mobile apps [3, 154, 155]. However, there is no consensus on the criteria to measure the quality of m-health apps.

One issue with the existing mobile app rating scales is that none of them have been developed based on a theoretical framework for the evaluation of m-health apps. Without a theoretical framework, the content validity of these scales would be questionable. In other words, app users cannot be confident if the existing rating scales include all assessment criteria that are relevant to the quality of m-health apps. Besides, non-expert potential app users such as older adults and adults with mental health conditions were not involved in the development process of the existing scales. As a result, the existing scales do not provide ratings that reflect the true experience of non-expert users of m-health apps.

To avoid this pitfall, a framework for quality of m-health apps was developed based on current theoretical models for usability and quality evaluation of mobile apps. The theoretical models and frameworks that were used included Unified Theory of Acceptance and Use of Technology [92], Seniors Technology Acceptance Model [91], Nielsen's model of usability [33],

App synopsis [95], Health On the Net code [96], Canadian Framework for Assessment of Mental Health Apps [94], and Health Care Information and Management Systems Society (HIMSS) framework [97]. Next, the framework was reviewed and approved by several groups of potential users of m-health apps who participated in six focus groups.

The user-oriented framework that we developed in phase 1 indicated that mobile app quality is a composite construct. Unlike hypothetical constructs that have (theoretically a universe of) observable correlated manifestations, composite constructs are defined by a number of finite independent criteria. From a theoretical perspective, hypothetical variables are bets measured with scales, and composite variables are best measured with indices. While assumptions of homogeneity such as internal consistency and loading of the items on the primary factor are crucial for scales, none of these hold for indices. Instead, an index must include all casual indicators, otherwise it may underestimate the composite construct of interest.

The findings of this study indicated that in general, there is a consensus over the most important criteria for quality of m-health apps between different stakeholders such as older adults, adults with mental health condition, health care providers, and app developers. The top most important criteria according to the participants in this study were usefulness, ease of use, security, and privacy. Also, health care providers insisted on including the trustworthiness of the developers and source of information provided within m-health apps. Other important criteria were customizability, affordability, and functionality.

Following the focus groups, the investigator generated an item pool for the identified criteria. The initial plan of the investigator was to generate the item pool based on the suggestions made by the participants in the focus groups in the previous phase. However, none of the participants were able to suggest items to assess the identified quality criteria of mobile

health apps. As an alternative, the investigator used the scales developed in previous studies as a source to generate the item pool for the index. The investigator extracted 771 items from the previous scales for rating mobile apps. Next, the investigator mapped the extracted items to the identified quality criteria. The mapping process was reviewed and revised by three experts in the field of technology evaluation and scale development. Similar items were merged, and irrelevant ones were deleted. At the end, an item pool was generated with 74 items unevenly distributed under nine quality criteria.

4.3. Content Validation of the Item Pool

The investigator used a two-step procedure to validate the content of the item pool. In the first step, he used an online survey to reach a wide range of audience. Participants in the survey were from four different provinces of Canada and included all spectrum of potential m-health app users, including clinicians, adults with mental health condition, older adults, and technology developers. Participants were requested to rate the relevance of items to the given criteria using a 3-point scale (1- Not relevant, 2- Relevant and 3- Unable to assess relevance). The investigator quantified participant's judgments using an index for content validity (CVI) by calculating the proportion of participants who endorsed an item as relevant [105]. To avoid inflation of agreement due to chance, Lynn's criteria were followed: only items with CVI $\geq$ 0.8 were accepted as valid items [105].

In the next step, six users of m-health apps were invited to join a focus group to further shortlist the most relevant items within each criterion and suggest any final modifications. The rigorous two step content validation process brings confidence that the index content are relevant and adequately capture the criteria to assess quality of m-health apps.

Readers may question why procedures such as factor analysis were not used to identify the most relevant items for this index. To respond to this critique, one needs to consider that the quality of apps is a composite construct defined by nine causal indicators (e.g., ease of use, security, affordability, etc.). However, there is no logical reason to believe that these causal indicators are related to each other. For example, if an app is easy to use, it does not mean that it is necessarily secure or affordable. Since the items of the index are not expected to be correlated with each other, the use of indices of homogeneity such as coefficient alpha, item-total correlation, and factor analysis are not appropriate [101]. Instead, as with all other assessment tools that measure composite constructs, it is important to make sure that the index includes all relevant criteria that define the quality of m-health apps.

4.4. Distribution of the Scores

The total score of ARIA for each app averaged across all 12 raters ranged between 41 and 52. The range of distribution was even wider within each group of raters (Health care providers: 29.5 – 51.5; Older adults: 42.5 – 58; Adults with mental health condition: 42 – 55.5). The wide range of distribution of total scores indicates that ARIA is capable of discriminating between apps based on its defined quality criteria.

Distribution of the ratings for items collapsed across raters and apps suggests that some items may benefit from revisions to the wording of the items or the scaling anchors. For example, the responses to items 1B, 10B, 11B, and 12B are distributed at both ends of the scaling spectrum (i.e., raters have either agreed or disagreed with these items). This observation suggests that perhaps these items should be rated using a dichotomous scale (i.e., Yes – No). Moreover, the distributions of the items 3A and 4A were positively skewed, which mean most of the raters have either completely agreed or completely disagreed with these items. Item 3A asked

if the app included a statement on the risks associated with using the app. Item 4A asked if the app declares any conflicts of interest. The positive skewness of the responses to these items suggests that either the attributes of interest were not present in most of the apps in this study, or the raters were not able to identify the attribute of interest in the apps. Further studies with a more diverse sample of apps are needed to clarify the real reason behind such observation.

4.5. Reliability of the Scores

Three groups of potential users of mental health mobile apps were recruited to test the performance of ARIA. Moreover, 50% of participants in each group were recruited from iPhone users and the other 50% from the Android users to eliminate any biases due to the operating system. On average, participants in the health care providers group rated the apps selected for this study lower than the other groups of raters. There are two potential reasons for this. First, participants in the health care providers group rated apps based on a very specific profile for a mock patient. The characteristics of the mock patient may have been very different from the characteristics of the participants in the other groups. Another reason can be that participants in the health care providers group were stricter with items relevant to trustworthiness, security, and privacy, and as a result, their total scores were lower compared to other groups.

The generalizability study of the index items indicated a small variance proportion associated with raters within groups, which indicates raters within each group were consistent. On the other hand, the largest proportion of variance was associated with the items, confirming that the construct under study (i.e., m-health app quality) is a composite construct, caused by several independent indicators (i.e., items).

The generalizability study of total scores of ARIA for the health care providers group indicated that the scores provided by health care providers to rate the quality of m-health apps

were reliable (G=0.948). Moreover, the variance component associated with raters was very low, which indicates raters in the health care providers group were very consistent.

The analysis of variance of the ratings provided evidence that the ratings made by rater number 4 in the older adults group and rater number 4 in the adults with mental health condition group were not consistent with the ratings made by their peers. These two participants, for some reasons, may have misunderstood the task, and as a result, their ratings did not reflect their true evaluation of the quality of apps used in this study. The findings from the generalizability study also showed that the variance proportion associated with the object of measurement (i.e., index total scores) and the G coefficient increased after these raters were excluded from the analyses. The high G coefficient, after removing the inconsistent raters, indicates that the total scores generated by older adults and adults with mental health condition, were reliable.

The results of the G study on star ratings were very interesting. For the health care providers group, the results were consistent with the total index score, which indicated that star ratings of app quality provided by health care providers were highly reliable. However, the variance proportion of the interaction between raters and star ratings was considerably high among participants in the older adults group and adults with mental health condition group which indicated a significant impact of potential rater idiosyncrasies and other systematic influences on the star ratings. This finding is consistent with other studies that reported star ratings by consumers as an unreliable measure for app quality [25, 26].

4.6. Comparability of the Ratings with Other Scales

Currently, there is no other gold standard scale for the quality of m-health apps. Moreover, the only other scale that has been designed for the general public is U-MARS. Participants in all groups were asked to rate the quality of the sample of mental health apps

twice: once with ARIA and once with U-MARS. The total scores from ARIA were compared to total scores of U-MARS. The results indicated low to moderate positive correlation between the total scores of ARIA and the total scores of U-MARS. The results suggest that both scales may measure unique and different attributes of the same underlying construct. However, further studies are needed to confirm the validity of the scores by each of these assessment tools.

4.7. Limitations

While the investigator employed rigorous methods and strategies to create and test ARIA, the study faces several limitations. First, the diversity of the mobile apps used in reliability testing was relatively low. However, since the app raters who participated in the study were from the general public (including older adults and adults with mental health condition), the investigator could not risk exposing the participants to apps that were very poor in terms of quality or were potentially harmful. For this reason, the sample of apps used in this study were selected from the directory of mental health apps of Alberta health services (AHS) addiction and mental health research partnership program [28]. Although the quality of apps in this directory have not been evaluated, they have been screened by AHS to make sure they are not very poor or harmful. As a result, the apps on the directory are similar to each other in many criteria. In spite of this, the diversity of the scores by raters indicated that the index is capable to distinguish between the apps that were more or less similar.

Second, the study could benefit from a larger sample of apps. However, considering the burden that evaluating apps with two scales might have on older adults and adults with mental health condition, the investigator kept the number of apps for this study at a convenient limit. Even with 11 apps, three participants were not able to complete the ratings in the given time and asked for an extension to complete the ratings.

The study could also benefit from a more diverse group of raters such as health care providers from other disciplines, adults with other health conditions, and apps for other health care issues. However, considering the timelines for this project, the scope of this study was limited to mental health apps and relevant stakeholders.

Perhaps the most notable limitation of the study is that the results provide only preliminary evidence for the validity of the ratings. However, this is a common issue shared by all assessment tools for the quality of m-health apps. Further studies with a focus on how the rating scores by this index may predict or impact acceptance and use of apps by different users are warranted to investigate the validity and usefulness of the index scores.

4.8. Clinical implications and statement of contribution

ARIA can help health care providers and patients to rate the quality of mobile health applications and find the ones that best meet their goals and expectations. Unlike other rating scales such as U-MARS that only provide a general rating for the quality of mobile health apps, ARIA can provide an individually tailored rating based on the goals, needs, and expectation of the user.

Currently, clinicians and the general public use the information from unreliable sources such as word of mouth, star ratings, advertisements, and reviews from unverified users on app stores to select mobile health apps. ARIA provides a validated checklist of criteria for health care providers and the general public to use as cues that can help them make an informed decision when selecting mobile health applications.

The ARIA was created primarily for use by the consumers of m-health apps. However, it may also benefit app developers and researchers in the field of health care informatics. Currently, the best practice is to engage potential users from the early stages of the design and development

of m-health apps. While this is a beneficial method to identify potential design issues sooner than later, participants may not be able to test all aspects of an app, especially if they do not have extensive experience related to the evaluation of the quality of mobile apps.

A criteria-based assessment tool, such as ARIA, can help m-health app developers in two ways. First, the index can educate app developers about the quality their m-health apps should meet to be acceptable by the general public. Second, app developers can provide ARIA to the participants in usability studies in order to use the quality criteria as cues to more comprehensively evaluate the quality of the app under development.

Although only mental health apps were used for reliability testing in this study, patients and clinicians can use ARIA to rate the quality of other health-related apps since the items in ARIA were designed to rate attributes of mobile health apps in general.

ARIA has a care provider version. Health care providers and family caregivers can use this version, as a proxy, to rate the quality of health apps for their clients or family members. ARIA has a user's version that can be used by the general public, including patients and older adults, to rate the quality of m-health apps for their use. We have also translated ARIA to French and will test reliability of the French version in the future studies (Appendix F and G).

The investigator strongly advises users to exercise caution when using any rating scales to rate the quality of m-health applications. Currently, most mobile health apps (unless specified otherwise) are considered consumer products, and therefore, health care providers and patients may only use them at their own risk. Use of quality rating scales such as those introduced in this study may help users make an informed decision and choose apps with better quality. However, no rating scale can identify all issues and risks associated with m-health apps. Moreover, rating

scales do not measure the effectiveness or validity of mobile health apps. Such attributes can only be determined through studies with focus on validity and clinical effectiveness.

As we mentioned earlier, mobile app quality is a composite construct, hence it is best measured with an index that includes all important quality criteria. None of the other mobile app rating scales are supported by frameworks of app quality evaluation or models of technology acceptance and use. As a result their ratings fall short of measuring all criteria that are important for measuring quality of mobile health apps. ARIA is grounded in frameworks of app evaluation and theories of technology acceptance and use. The content of ARIA has also been validated by users of mobile health apps as well as health informatics experts, which gives this index an advantage in comparison to the other scales.

4.9. Future Research

The results of this study provided a guideline for future studies to evaluate the quality of information technologies, including mobile health apps. ARIA can be further validated on apps relevant to different health care conditions and by different consumers, including health care providers in different disciplines and patients with health conditions other than mental health issues. App developers and researchers can also use ARIA to develop health-related apps to better address the needs and goals of patients, health care providers, and family caregivers.

One important step is to disseminate ARIA and make it accessible to potential users. The investigator intends to develop and validate an online (web-based) version of ARIA for the general public. App users who are interested in rating a mobile health app can use this tool instead of the paper-based version. Users can log in to the website and answer relevant questions about their experience using the app. The web-based ARIA will automatically report the final total quality score as well as a short recommendation about using the app under evaluation. The

web-based version can also record the ratings provided for each app and eventually create a crowd-sourced directory for mobile health applications. We have also translated ARIA into French to make it more accessible to Canadians. We will test the reliability of the French version in the web-based form in future studies.

4.10. Conclusion

The number of mobile health applications is rapidly increasing. However, minimal regulations and control mechanisms on online app markets have increased the risk of the public using apps with poor quality or apps that are harmful. Previous attempts to create a scale to evaluate the quality of health apps have been focused on expert consumers. Besides, none of these scales were grounded in theoretical frameworks of usability assessment and quality evaluation of mobile apps.

In this study, an index was developed based on theories of technology acceptance as well as frameworks for evaluation of the quality of mobile health apps. The content of the index was rigorously validated to make sure it includes all relevant criteria and items to measure the quality of m-health apps. The index was named Alberta Rating Index for Apps (ARIA).

The study confirmed that at least nine criteria rate the quality of mobile health apps. These criteria are purpose, trustworthiness, privacy, security, affordability, ease of use, functionality, appropriateness for target users, and usefulness and satisfaction.

The results from the reliability study confirmed that health care providers, older adults, and adults with mental health condition may reliably use ARIA to rate the quality of m-health apps according to their goals, needs, and preferences. Moreover, the users of m-health apps prefer ARIA in comparison to U-MARS since it is more comprehensive and easier to use. Based

on these findings, it is concluded that ARIA is superior to other app quality rating scales because of its content validity, reliability, and ease of use.

REFERENCES

1. Lui, J.H.M., David K.; Barry, Christopher T., *Evidence-based apps? A review of mental health mobile applications in a psychotherapy context.* Professional Psychology: Research and Practice, 2017. 48(3): p. 199-210.
2. BinDhim, N.F., A. Hawkey, and L. Trevena, *A systematic review of quality assessment methods for smartphone health apps.* Telemedicine Journal And E-Health: The Official Journal Of The American Telemedicine Association, 2015. 21(2): p. 97-104.
3. Grundy, Q.H., Z. Wang, and L.A. Bero, *Challenges in Assessing Mobile Health App Quality: A Systematic Review of Prevalent and Innovative Methods.* American Journal Of Preventive Medicine, 2016. 51(6): p. 1051-1059.
4. Merrick, L. Different types of apps that app developers need to know about. 2016 [cited 29/02/2020]; Available from: https://www.buzinga.com.au/buzz/how-many-app-types-are-there/.
5. Stoyanov, S.R., et al., *Mobile app rating scale: a new tool for assessing the quality of health mobile apps.* JMIR Mhealth Uhealth, 2015. 3(1).
6. Boxall, A. *2014 is the year of health and fitness apps, says Google*. 2014 05/08/2016]; Available from: http://www.digitaltrends.com/mobile/google-play-store-2014-most-downloaded-apps/.
7. Research2Guidance, *mHealth App Economics 2017/2018*. 2017.
8. IMS institute of health informatics, *New report finds more than 165,000 mobile health apps now available, takes close look at characteristics & use*. 2015 [cited 2018 25/04]; Available from: https://www.imedicalapps.com/2015/09/ims-health-apps-report/#.

9. Donker, T., et al., *Smartphones for smarter delivery of mental health programs: a systematic review.* J Med Internet Res, 2013. 15(11).

10. National Institute of Mental Health (NIMH), *Technology and the Future of Mental Health Treatment*. 2017 [cited 2018 25/04]; Available from: https://www.nimh.nih.gov/health/topics/technology-and-the-future-of-mental-health-treatment/index.shtml.

11. Larsen, M.E., J. Nicholas, and H. Christensen, *Quantifying App Store Dynamics: Longitudinal Tracking of Mental Health Apps.* JMIR Mhealth And Uhealth, 2016. 4(3): p. e96-e96.

12. Lui, J.H., D.K. Marcus, and C.T. Barry, *Evidence-based apps? A review of mental health mobile applications in a psychotherapy context.* Professional Psychology: Research and Practice, 2017. 48(3): p. 199-210.

13. Nicholas, J., et al., *Mobile Apps for Bipolar Disorder: A Systematic Review of Features and Content Quality.* J Med Internet Res, 2015. 17(8): p. e198.

14. Bindhim, N.F., et al., *Apps Promoting Illicit Drugs-A Need for Tighter Regulation?* Journal of Consumer Health on the Internet, 2014. 18(1): p. 31-43.

15. Leigh, S., *Comparing applets and oranges: barriers to evidence-based practice for app-based psychological interventions.* Evidence Based Mental Health, 2016. 19(3): p. 90-92.

16. Smith, A., *Older Adults and Technology Use*. 2014, Pew Research Center. Available from: https://www.pewresearch.org/internet/2014/04/03/older-adults-and-technology-use/.

17. Anderson, G.O., *Getting Connected: Older Americans Embrace Technology to Enhance Their Lives.* 2017, AARP Research: Washington, DC. Available from:

https://www.aarp.org/research/topics/technology/info-2018/technology-use-attitudes.html.

18. Wildenbos, G.A., L.W. Peute, and M.W. Jaspers, *A framework for evaluating mHealth tools for Older Patients on Usability.* Stud Health Technol Inform, 2015. 210: p. 783-7.

19. Food and Drug Administration (FDA), *Mobile Medical Applications: Guidance for Industry and Food and Drug Administration Staff*. 2015; Available from: https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM263366.pdf.

20. Canada, *Medical Devices Regulations SOR/98-282*. 2018. Available from: https://laws-lois.justice.gc.ca/eng/regulations/sor-98-282/.

21. Apple. *App Review*. 2018 [cited 2018 20/05]; Available from: https://developer.apple.com/support/app-review/.

22. Google, *Android Developer Quality Guidelines: Core App Quality*. 2018. Available from: https://developer.android.com/docs/quality-guidelines/core-app-quality.

23. Armstrong, S., *Which app should I use?* BMJ: British Medical Journal, 2015. 351(8027): p. h4597-2.

24. BinDhim, N.F.H., Alexandra; Trevena, Lyndal, *A systematic review of quality assessment methods for smartphone health apps.* Telemedicine Journal And E-Health: The Official Journal Of The American Telemedicine Association, 2015. 21(2): p. 97-104.

25. Fiore, P., *How to Evaluate Mobile Health Applications: A Scoping Review.* Studies In Health Technology And Informatics, 2017. 234: p. 109-114.

26. Singh, K., et al., *Many Mobile Health Apps Target High-Need, High-Cost Populations, But Gaps Remain.* Health Affairs, 2016. 35(12): p. 2310-2318.

27. National Health Servides (NHS), *NHS Apps Library (beta)*. 2017 [cited 2018 22/05]; Available from: https://digital.nhs.uk/services/nhs-apps-library.

28. Alberta Addiction & Mental Health Research Partnership Program. (2019). Addiction and Mental Health - Mobile Application Directory 2019. Edmonton, AB: Alberta Health Services. Available from: https://www.albertahealthservices.ca/assets/info/res/mhr/if-res-mhr-kt-mobile-app-directory.pdf.

29. Women's College Hospital Institute for Health System Solutions and Virtual Care and Ontario Telemedicine Network, *practicalapps*. 2016; Available from: https://practicalapps.ca/.

30. Women's College Hospital Institute for Health System Solutions and Virtual Care and Ontario Telemedicine Network, *Practicalapps: our approach*. 2016; Available from: https://practicalapps.ca/approach/

31. Larsen, M.E.N., Jennifer; Christensen, Helen, *Quantifying App Store Dynamics: Longitudinal Tracking of Mental Health Apps.* JMIR Mhealth And Uhealth, 2016. 4(3): p. e96-e96.

32. Wicks, P. and E. Chiauzzi, *"Trust but verify"- five approaches to ensure safe medical apps.* BMC medicine, 2015. 13: p. 205-205.

33. Nielsen, J., *Usability engineering. [electronic resource]*. 1993: Cambridge, Mass. : AP Professional, c1993.

34. Brooke, J., *SUS: A "quick and dirty" usability scale*, in *Usability Evaluation in Industry.*, T.B. Jordan PW, Weerdmeester BA, and McClelland AL, Editors. 1996, Taylor and Francis: London. p. 189–94.

35. Alnasser, A., et al., *Relationship Between Evidence Requirements, User Expectations, and Actual Experiences: Usability Evaluation of the Twazon Arabic Weight Loss App.* JMIR Human Factors, 2018. 5(2): p. e16-e16.

36. Ehrler, F., et al., *Acceptance of a Mobile Application Supporting Nurses Workflow at Patient Bedside: Results from a Pilot Study.* Studies In Health Technology And Informatics, 2018. 247: p. 506-510.

37. Ginsburg, A.S., et al., *mPneumonia, an Innovation for Diagnosing and Treating Childhood Pneumonia in Low-Resource Settings: A Feasibility, Usability and Acceptability Study in Ghana.* Plos One, 2016. 11(10): p. e0165201-e0165201.

38. Isaković, M., et al., *Usability Pitfalls of Diabetes mHealth Apps for the Elderly.* Journal of Diabetes Research, 2016: p. 1-9.

39. Ithnin, M., et al., *Mobile App Design, Development, and Publication for Adverse Drug Reaction Assessments of Causality, Severity, and Preventability.* JMIR Mhealth And Uhealth, 2017. 5(5): p. e78-e78.

40. Davis, F.D., *Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology.* MIS Quarterly, 1989. 13(3): p. 319-340.

41. Price, M.S., Tyler; Harris, Madison; Skalka, Christian, *Usability Evaluation of a Mobile Monitoring System to Assess Symptoms After a Traumatic Injury: A Mixed-Methods Study.* JMIR Mental Health, 2016. 3(1): p. e3-e3.

42. O'Malley, G., et al., *Exploring the usability of a mobile app for adolescent obesity management.* Physiotherapy, 2016. 102: p. e44-e45.

43. Bondaronek, P., et al., *Quality of Publicly Available Physical Activity Apps: Review and Content Analysis.* JMIR Mhealth And Uhealth, 2018. 6(3): p. e53-e53.

44. Brooks, G.C., et al., *Accuracy and Usability of a Self-Administered 6-Minute Walk Test Smartphone Application.* Circulation. Heart Failure, 2015. 8(5): p. 905-913.

45. Ehrler, F., et al., *A Mobile App (BEDSide Mobility) to Support Nurses' Tasks at the Patient's Bedside: Usability Study.* JMIR Mhealth And Uhealth, 2018. 6(3): p. e57-e57.

46. Fallah, M. and M. Yasini, *A Medication Reminder Mobile App: Does It Work for Different Age Ranges.* Studies In Health Technology And Informatics, 2017. 235: p. 68-72.

47. Ginsburg, A.S., et al., *mPneumonia: Development of an Innovative mHealth Application for Diagnosing and Treating Childhood Pneumonia and Other Childhood Illnesses in Low-Resource Settings.* Plos One, 2015. 10(10): p. e0139625-e0139625.

48. Grindrod, K., et al., *Evaluating authentication options for mobile health applications in younger and older adults.* Plos One, 2018. 13(1): p. e0189048-e0189048.

49. Mattson, D.C., *Usability assessment of a mobile app for art therapy.* The Arts in Psychotherapy, 2015. 43: p. 1-6.

50. Nitsch, M., et al., *A Guided Online and Mobile Self-Help Program for Individuals With Eating Disorders: An Iterative Engagement and Usability Study.* Journal Of Medical Internet Research, 2016. 18(1): p. e7-e7.

51. Sullivan, P.S., et al., *Usability and Acceptability of a Mobile Comprehensive HIV Prevention App for Men Who Have Sex With Men: A Pilot Study.* JMIR Mhealth And Uhealth, 2017. 5(3): p. e26-e26.

52. Villalobos, O., et al., *Utilization of a Mobile App to Assess Psychiatric Patients With Limited English Proficiency.* Hispanic Journal of Behavioral Sciences, 2017. 39(3): p. 369-380.

53. Mattson, D.C., *Usability evaluation of the digital anger thermometer app.* Health Informatics Journal, 2016.

54. Barrio, P., et al., *Self-management and shared decision-making in alcohol dependence via a mobile app: A pilot study.* International Journal of Behavioral Medicine, 2017: p. No Pagination Specified.

55. Ward, G., A. Walker-Clarke, and N. Holliday, *Evaluation of a web-based app to assist home-hazard modification in falls prevention.* British Journal of Occupational Therapy, 2017. 80(12): p. 735-744.

56. Vélez, O., et al., *A usability study of a mobile health application for rural Ghanaian midwives.* Journal Of Midwifery & Women's Health, 2014. 59(2): p. 184-191.

57. Schnall, R., H. Cho, and J. Liu, *Health Information Technology Usability Evaluation Scale (Health-ITUES) for Usability Assessment of Mobile Health Technology: Validation Study.* JMIR Mhealth And Uhealth, 2018. 6(1): p. e4-e4.

58. Stonbraker, S., et al., *Usability Testing of a mHealth App to Support Self-Management of HIV-Associated Non-AIDS Related Symptoms.* Studies in Health Technology & Informatics, 2018. 250: p. 106-110.

59. English, L.L., et al., *The PAediatric Risk Assessment (PARA) Mobile App to Reduce Postdischarge Child Mortality: Design, Usability, and Feasibility for Health Care Workers in Uganda.* JMIR Mhealth And Uhealth, 2016. 4(1): p. e16-e16.

60. Torous, J.B., et al., *To Use or Not? Evaluating ASPECTS of Smartphone Apps and Mobile Technology for Clinical Care in Psychiatry.* The Journal Of Clinical Psychiatry, 2016. 77(6): p. e734-e738.

61. Chyjek, K., S. Farag, and K.T. Chen, *Rating Pregnancy Wheel Applications Using the APPLICATIONS Scoring System.* Obstetrics And Gynecology, 2015. 125(6): p. 1478-1483.

62. Perry, R., B. Lunde, and K.T. Chen, *An evaluation of contraception mobile applications for providers of family planning services.* Contraception, 2016. 93(6): p. 539-544.

63. Loy, J.S., E.E. Ali, and K.Y.-L. Yap, *Quality Assessment of Medical Apps that Target Medication-Related Problems.* Journal Of Managed Care & Specialty Pharmacy, 2016. 22(10): p. 1124-1140.

64. Baumel, A., et al., *Enlight: A Comprehensive Quality and Therapeutic Potential Evaluation Tool for Mobile and Web-Based eHealth Interventions.* Journal Of Medical Internet Research, 2017. 19(3): p. e82-e82.

65. DiFilippo, K.N., W. Huang, and K.M. Chapman-Novakofski, *A New Tool for Nutrition App Quality Evaluation (AQEL): Development, Validation, and Reliability Testing.* JMIR Mhealth And Uhealth, 2017. 5(10): p. e163-e163.

66. Ali, E.E., et al., *MedAd-AppQ: A quality assessment tool for medication adherence apps on iOS and android platforms.* Research In Social & Administrative Pharmacy: RSAP, 2018.

67. Anderson, K., O. Burford, and L. Emmerton, *App Chronic Disease Checklist: Protocol to Evaluate Mobile Apps for Chronic Disease Self-Management.* JMIR Research Protocols, 2016. 5(4): p. e204-e204.

68. Jin, M. and J. Kim, *Development and Evaluation of an Evaluation Tool for Healthcare Smartphone Applications.* Telemedicine Journal And E-Health: The Official Journal Of The American Telemedicine Association, 2015. 21(10): p. 831-837.

69. Wyatt, J.C., et al., *What makes a good clinical app? Introducing the RCP Health Informatics Unit checklist.* Clinical Medicine (London, England), 2015. 15(6): p. 519-521.

70. Leigh, S., J. Ouyang, and C. Mimnagh, *Effective? Engaging? Secure? Applying the ORCHA-24 framework to evaluate apps for chronic insomnia disorder.* Evidence Based Mental Health, 2017. 20(4): p. e20-e20.

71. Mani, M., et al., *Review and Evaluation of Mindfulness-Based iPhone Apps.* JMIR Mhealth Uhealth, 2015. 3(3).

72. Bardus, M., et al., *A review and content analysis of engagement, functionality, aesthetics, information quality, and change techniques in the most popular commercial apps for weight management.* The International Journal Of Behavioral Nutrition And Physical Activity, 2016. 13: p. 35-35.

73. Furlong, L.M., et al., *Quality of Mobile Phone and Tablet Mobile Apps for Speech Sound Disorders: Protocol for an Evidence-Based Appraisal.* JMIR Research Protocols, 2016. 5(4): p. e233-e233.

74. Hartin, P.J., et al., *Assessing app quality through expert peer review: a case study from the gray matters study.* Conference Proceedings: ... Annual International Conference Of The IEEE Engineering In Medicine And Biology Society. IEEE Engineering In Medicine And Biology Society. Annual Conference, 2016. 2016: p. 4379-4382.

75. Masterson Creber, R.M., et al., *Review and Analysis of Existing Mobile Phone Apps to Support Heart Failure Symptom Monitoring and Self-Care Management Using the Mobile Application Rating Scale (MARS).* JMIR Mhealth And Uhealth, 2016. 4(2): p. e74-e74.

76. Reyes, A., P. Qin, and C.A. Brown, *A standardized review of smartphone applications to promote balance for older adults.* Disability And Rehabilitation, 2016: p. 1-10.

77. Böhme, C., et al., *Development of a Rating Tool for Mobile Cancer Apps: Information Analysis and Formal and Content-Related Evaluation of Selected Cancer Apps.* Journal Of Cancer Education: The Official Journal Of The American Association For Cancer Education, 2017.

78. Böhme, C., et al., *Qualitative evaluation of mobile cancer apps with particular attention to the target group, content, and advertising.* Journal Of Cancer Research And Clinical Oncology, 2018. 144(1): p. 173-181.

79. Escoffery, C., et al., *A review of mobile apps for epilepsy self-management.* Epilepsy & Behavior: E&B, 2018. 81: p. 62-69.

80. Flaherty, S.-J., et al., *Can existing mobile apps support healthier food purchasing behaviour? Content analysis of nutrition content, behaviour change theory and user quality integration.* Public Health Nutrition, 2018. 21(2): p. 288-298.

81. Fraynt, R., et al., *An evaluation of mobile applications designed to assist service members and veterans transitioning to civilian life.* Special Issue: Psychosocial Approaches to Improving the Military-to-Civilian Transition Process, 2018. 15(2): p. 208-215.

82. Jisan, L. and K. Jeongeun, *Method of App Selection for Healthcare Providers Based on Consumer Needs.* CIN: Computers, Informatics, Nursing, 2018. 36(1): p. 45-54.

83. Jupp, J.C.Y., et al., *Evaluation of mobile phone applications to support medication adherence and symptom management in oncology patients.* Pediatric Blood & Cancer, 2018: p. e27278-e27278.

84. Kim, B.Y., et al., *Consumer Mobile Apps for Potential Drug-Drug Interaction Check: Systematic Review and Content Analysis Using the Mobile App Rating Scale (MARS).* JMIR Mhealth And Uhealth, 2018. 6(3): p. e74-e74.

85. Le Marne, F.A., et al., *EpApp: Development and evaluation of a smartphone/tablet app for adolescents with epilepsy.* Journal Of Clinical Neuroscience: Official Journal Of The Neurosurgical Society Of Australasia, 2018. 50: p. 214-220.

86. Tinschert, P., et al., *The Potential of Mobile Apps for Improving Asthma Self-Management: A Review of Publicly Available and Well-Adopted Asthma Apps.* JMIR Mhealth And Uhealth, 2017. 5(8): p. e113-e113.

87. Stoyanov, S.R., et al., *Development and Validation of the User Version of the Mobile Application Rating Scale (uMARS).* JMIR Mhealth And Uhealth, 2016. 4(2): p. e72-e72.

88. Davis, F.D., R.P. Bagozzi, and P.R. Warshaw, *User Acceptance of Computer Technology: A Comparison of Two Theoretical Models.* Management Science, 1989. 35(8): p. 982-1003.

89. Rogers, E.M., *Diffusion of innovations*. 2003: New York : Free Press, 2003.

5th ed., Free Press trade pbk. ed.

90. Venkatesh, V., J.Y.L. Thong, and X. Xu, *Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology.* MIS Quarterly, 2012(1): p. 157.

91. Renaud, K. and J.v. Biljon, *Predicting technology acceptance and adoption by the elderly: a qualitative study*, in *Proceedings of the 2008 annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT*

research in developing countries: riding the wave of technology. 2008, ACM: Wilderness, South Africa. p. 210-219.

92. Venkatesh, V., et al., *User Acceptance of Information Technology: Toward a Unified View.* MIS Quarterly, 2003. 27(3): p. 425-478.

93. Nielsen, J. *Usability 101: Introduction to usability* 2014 [cited 2016 11/09]; Available from: https://www.nngroup.com/articles/usability-101-introduction-to-usability/.

94. Zelmer, J., et al., *An assessment framework for e-mental health apps in Canada: Results of a modified delphi process.* Journal of Medical Internet Research, 2018. 20(7).

95. UrsVito Albrecht; Pramann O, v.J.U. *App-Synopsis - Standard Reporting for Health-Apps*. 2013 [cited 2018 25/05]; Available from: http://plrimedapplab.weebly.com/app-synopsis.html.

96. Health on the net (HON). *The commitment to reliable health and medical information on the internet* 2017 [cited 2019 18/01]; Available from: https://www.hon.ch/HONcode/Visitor/visitor.html.

97. Chan, S., et al., *Towards a Framework for Evaluating Mobile Mental Health Apps.* Telemed J E Health, 2015. 14: p. 14.

98. Ben-Mussa, A. and A.M. Paget, *Popular apps on the medical category targeting patients and the general public in the United Kingdom: Do they conform to the Health On the Net Foundation principles?* Health Informatics Journal, 2016.

99. Health on the net (HON), *About Health On the Net*. 2018 [cited 2018 25/05/2018]; Available from: https://www.healthonnet.org/en/about.html.

100. Mental health commission of Canada, *Mental Health Apps: How to Make an Informed Choice*. 2016; Available from: https://www.mentalhealthcommission.ca/sites/default/files/2018-01/eMH_app_eng.pdf.

101. Streiner, D.L., Norman, Geoffrey R., Cairney, John, *Health measurement scales : a practical guide to their development and use*, G.R. Norman and J. Cairney, Editors. 2015.

102. Patton, M.Q., *Qualitative evaluation and research methods*. Vol. 2nd. 1990, Newbury Park, CA: Sage Publications.

103. Braun, V. and V. Clarke, *Using thematic analysis in psychology.* Qualitative Research in Psychology, 2006. 3(2): p. 77-101.

104. Nouri, R., et al., *Criteria for assessing the quality of mHealth apps: a systematic review.* Journal of the American Medical Informatics Association, 2018: p. ocy050-ocy050.

105. Lynn, M.R., *Determination and quantification of content validity.* Nursing Research, 1986. 35(6): p. 382-386.

106. Brennan, R.L., *Generalizability theory*. 2001, Springer: New York :.

107. Bennett, S.J., et al., *Discriminant properties of commonly used quality of life measures in heart failure.* Quality of Life Research, 2002. 11(4): p. 349-359.

108. *IBM SPSS Statistics for Windows*. 2017, IBM Corp: Armonk, NY.

109. Mushquash, C. and B.P. O'Connor, *SPSS and SAS programs for generalizability theory analyses.* Behavior Research Methods, 2006. 38(3): p. 542-547.

110. Moher, D., et al., *Preferred reportng items for systematic reviews and meta-analyses: the PRISMA statement.* Int J Surg, 2010. 8(5): p. 336-41.

111. Price, M., et al., *Usability Evaluation of a Mobile Monitoring System to Assess Symptoms After a Traumatic Injury: A Mixed-Methods Study.* JMIR Mental Health, 2016. 3(1): p. e3-e3.

112. Lewis, J.R., *IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use.* International Journal of Human–Computer Interaction, 1995. 7(1): p. 57-78.

113. Gediga, G., K.-C. Hamborg, and I. Düntsch, *The IsoMetrics usability inventory: An operationalization of ISO 9241-10 supporting summative and formative evaluation of software systems.* Behaviour & Information Technology, 1999. 18(3): p. 151-164.

114. Lund, A.M., *Measuring usability with the USE questionnaire*, in *Usability Interface*. 2001. p. 3-6.

115. Yen, P.-Y., D. Wantland, and S. Bakken, *Development of a Customizable Health IT Usability Evaluation Scale.* AMIA ... Annual Symposium proceedings. AMIA Symposium, 2010. 2010: p. 917-921.

116. Martinez-Perez, B., et al., *Development and evaluation of tools for measuring the quality of experience (QoE) in mHealth applications.* J Med Syst, 2013. 37(5): p. 013-9976.

117. McMillan, B., et al., *Quality assessment of a sample of mobile app-based health behavior change interventions using a tool based on the National Institute of Health and Care Excellence behavior change guidance.* Patient Education And Counseling, 2016. 99(3): p. 429-435.

118. Guo, Y., et al., *Assessing the Quality of Mobile Exercise Apps Based on the American College of Sports Medicine Guidelines: A Reliable and Valid Scoring Instrument.* Journal Of Medical Internet Research, 2017. 19(3): p. e67-e67.

119. Van Singer, M., A. Chatton, and Y. Khazaal, *Quality of Smartphone Apps Related to Panic Disorder.* Frontiers in psychiatry, 2015. 6: p. 96-96.

120. Gibbs, J., et al., *'Can you recommend any good STI apps?' A review of content, accuracy and comprehensiveness of current mobile medical applications for STIs and related genital infections.* Sexually Transmitted Infections, 2017. 93(4): p. 234-235.

121. Huckvale, K., et al., *Apps for asthma self-management: a systematic assessment of content and tools.* BMC Medicine, 2012. 10(1): p. 144.

122. Househ, M., et al., *A cross-sectional content analysis of Android applications for asthma.* Health Informatics Journal, 2017. 23(2): p. 83-95.

123. Martínez-Pérez, B., I. de la Torre-Díez, and M. López-Coronado, *Experiences and Results of Applying Tools for Assessing the Quality of a mHealth App Named Heartkeeper.* Journal Of Medical Systems, 2015. 39(11): p. 142-142.

124. Jeon, E., et al., *Analysis of the information quality of korean obesity-management smartphone applications.* Healthcare informatics research, 2014. 20(1): p. 23-29.

125. Zhang, M.W.B., et al., *Analysis of the Information Quality of Bariatric Surgery Smartphone Applications Using the Silberg Scale.* Obesity Surgery, 2016. 26(1): p. 163-168.

126. Zhang, M.W., et al., *Current status of postnatal depression smartphone applications available on application stores: an information quality analysis.* BMJ Open, 2017. 7(11): p. e015655-e015655.

127. Spook, J.E., et al., *Monitoring dietary intake and physical activity electronically: feasibility, usability, and ecological validity of a mobile-based Ecological Momentary Assessment tool.* Journal Of Medical Internet Research, 2013. 15(9): p. e214-e214.

128. Arnhold, M., M. Quade, and W. Kirch, *Mobile Applications for Diabetics: A Systematic Review and Expert-Based Usability Evaluation Considering the Special Requirements of Diabetes Patients Age 50 Years or Older.* JOURNAL OF MEDICAL INTERNET RESEARCH, 2014. 16(4): p. 18.

129. Chenchen, G., et al., *Mobile application for diabetes self-management in China: Do they fit for older adults?* International Journal of Medical Informatics, 2017. 101: p. 68-74.

130. O'Connor, E., M. Farrow, and C. Hatherly, *Randomized Comparison of Mobile and Web-Tools to Provide Dementia Risk Reduction Education: Use, Engagement and Participant Satisfaction.* JMIR Mental Health, 2014. 1(1): p. e4-e4.

131. Al Ayubi, S.U., et al., *A Persuasive and Social mHealth Application for Physical Activity: A Usability and Feasibility Study.* JMIR Mhealth And Uhealth, 2014. 2(2): p. e25-e25.

132. Kuhn, E., et al., *Preliminary evaluation of PTSD Coach, a smartphone app for post-traumatic stress symptoms.* Mil Med, 2014. 179(1): p. 12-8.

133. Cho, M.J., J.L. Sim, and S.Y. Hwang, *Development of smartphone educational application for patients with coronary artery disease.* Healthcare Informatics Research, 2014. 20(2): p. 117-124.

134. Rizvi, S.L., C.D. Hughes, and M.C. Thomas, *The DBT Coach mobile application as an adjunct to treatment for suicidal and self-injuring individuals with borderline personality disorder: A preliminary evaluation and challenges to client utilization.* Psychological Services, 2016. 13(4): p. 380-388.

135. Wilson, K., et al., *An evaluation of the feasibility and usability of a proof of concept mobile app for adverse event reporting post influenza vaccination.* Human Vaccines & Immunotherapeutics, 2016. 12(7): p. 1738-1748.

136. Tay, I., et al., *Development and Testing of a Mobile Phone App for Self-Monitoring of Calcium Intake in Young Women.* JMIR Mhealth And Uhealth, 2017. 5(3): p. e27-e27.

137. de la Torre Díez, I., et al., *A New mHealth App for Monitoring and Awareness of Healthy Eating: Development and User Evaluation by Spanish Users.* Journal Of Medical Systems, 2017. 41(7): p. 109-109.

138. Zapata, B.C., et al., *Assessing the privacy policies in mobile personal health records.* Conference Proceedings: ... Annual International Conference Of The IEEE Engineering In Medicine And Biology Society. IEEE Engineering In Medicine And Biology Society. Annual Conference, 2014. 2014: p. 4956-4959.

139. Ouhbi, S., et al., *Compliance of blood donation apps with mobile OS usability guidelines.* Journal Of Medical Systems, 2015. 39(6): p. 63-63.

140. Cruz Zapata, B., et al., *Mobile PHRs compliance with Android and iOS usability guidelines.* Journal Of Medical Systems, 2014. 38(8): p. 81-81.

141. Ramos, S.R., *User-Centered Design, Experience, and Usability of an Electronic Consent User Interface to Facilitate Informed Decision-Making in an HIV Clinic.* CIN: Computers, Informatics, Nursing, 2017. 35(11): p. 556-564.

142. Myint, M., et al., *Mobile phone applications in management of enuresis: The good, the bad, and the unreliable!* Journal Of Pediatric Urology, 2016. 12(2): p. 112.e1-6.

143. Fiks, A.G., et al., *Usability, Acceptability, and Impact of a Pediatric Teledermatology Mobile Health Application.* Telemedicine Journal And E-Health: The Official Journal Of The American Telemedicine Association, 2018. 24(3): p. 236-245.

144. Robustillo Cortés, M.d.l.A.C.C., María Rosa; Morillo Verdugo, Ramón; Calvo Cidoncha, Elena, *High quantity but limited quality in healthcare applications intended for HIV-infected patients.* Telemedicine Journal And E-Health: The Official Journal Of The American Telemedicine Association, 2014. 20(8): p. 729-735.

145. de la Vega, R., et al., *Fibroline: A mobile app for improving the quality of life of young people with fibromyalgia.* Journal of Health Psychology, 2018. 23(1): p. 67-78.

146. Reynoldson, C., et al., *Assessing the quality and usability of smartphone apps for pain self-management.* Pain Medicine (Malden, Mass.), 2014. 15(6): p. 898-909.

147. Akinola, M., et al., *Development of a Mobile App on Contraceptive Options for Young African American and Latina Women.* Health Education & Behavior: The Official Publication Of The Society For Public Health Education, 2018: p. 1090198118775476-1090198118775476.

148. Butcher, R., et al., *Development and examination of a rubric for evaluating point-of-care medical applications for mobile devices.* Medical Reference Services Quarterly, 2015. 34(1): p. 75-87.

149. Wang, C.-J., P. Chaovalit, and S. Pongnumkul, *A Breastfeed-Promoting Mobile App Intervention: Usability and Usefulness Study.* JMIR Mhealth And Uhealth, 2018. 6(1): p. e27-e27.

150. Choi, J., G.-Y. Noh, and D.-J. Park, *Smoking cessation apps for smartphones: content analysis with the self-determination theory.* Journal Of Medical Internet Research, 2014. 16(2): p. e44-e44.

151. Burdette, S.D., R. Trotman, and J. Cmar, *Mobile infectious disease references: from the bedside to the beach.* Clinical Infectious Diseases: An Official Publication Of The Infectious Diseases Society Of America, 2012. 55(1): p. 114-125.

152. Huguet, A., et al., *A Systematic Review of Cognitive Behavioral Therapy and Behavioral Activation Apps for Depression.* Plos One, 2016. 11(5): p. e0154248-e0154248.

153. Roduta Roberts, M. and C.M. Gotch, *Development and Examination of a Tool to Assess Score Report Quality.* Frontiers in Education, 2019. 4(20).

154. Abbott, V.P., *Web page quality: can we measure it and what do we find? A report of exploratory findings.* Journal of Public Health, 2000. 22(2): p. 191-197.

155. Mendiola, M.F., M. Kalnicki, and S. Lindenauer, *Valuable features in mobile health apps for patients and consumers: content analysis of apps and user ratings.* JMIR Mhealth Uhealth, 2015. 3(2).

156. Zapata, B.C., et al., *Empirical Studies on Usability of mHealth Apps: A Systematic Literature Review*.

157. Zhang, M.W.B., et al., *Pilot implementation and user preferences of a Bariatric After-care application.* Technology And Health Care: Official Journal Of The European Society For Engineering And Medicine, 2015. 23(6): p. 729-736.

158. DiFilippo, K.N., et al., *The use of mobile apps to improve nutrition outcomes: A systematic literature review.* Journal Of Telemedicine And Telecare, 2015. 21(5): p. 243-253.

159. Barrio, P., et al., *Development, Validation, and Implementation of an Innovative Mobile App for Alcohol Dependence Management: Protocol for the SIDEAL Trial.* JMIR Research Protocols, 2016. 5(1): p. e27-e27.

160. Grindrod, K.A., et al., *ClereMed: Lessons Learned From a Pilot Study of a Mobile Screening Tool to Identify and Support Adults Who Have Difficulty With Medication Labels.* JMIR Mhealth And Uhealth, 2014. 2(3): p. e35-e35.

161. Wilson, K., et al., *Improving vaccine registries through mobile technologies: a vision for mobile enhanced Immunization information systems.* Journal Of The American Medical Informatics Association: JAMIA, 2016. 23(1): p. 207-211.

162. McMillan, B., et al., *The need for quality assurance of health apps.* BMJ (Clinical Research Ed.), 2015. 351: p. h5915-h5915.

163. Atkinson, K.M., et al., *Can mobile technologies improve on-time vaccination? A study piloting maternal use of ImmunizeCA, a Pan-Canadian immunization app.* Human Vaccines & Immunotherapeutics, 2016. 12(10): p. 2654-2661.

164. Brooks, B. and C. Gustus, *HANDS ON: APPS AND THE IPAD FOR CHILDREN WITH HEARING LOSS.* Volta Voices, 2016. 23(1): p. 12-17.

APPENDICES

Appendix A: Notifications of Ethics Approvals

Date:	February 9, 2017	
Study ID:	Pro00067856	
Principal Investigator:	Peyman Azad Khaneghah	
Study Supervisor:	Lili Liu	
Study Title:	A Rating Scale for Mental Health Mobile Applications for Older Adults	
Approval Expiry Date:	Thursday, February 8, 2018	
Approved Consent Form:	Approval Date	Approved Document
	2/9/2017	phase 3 information letter and consent form
	2/9/2017	Phase 2 information letter and consent form
	2/9/2017	Phase 1 information letter and consent form
Sponsor/Funding Agency:	AGE-WELL NCE	

Thank you for submitting the above study to the Research Ethics Board 2. Your application has been reviewed and approved on behalf of the committee.

A renewal report must be submitted next year prior to the expiry of this approval if your study still requires ethics approval. If you do not renew on or before the renewal expiry date, you will have to re-submit an ethics application.

Approval by the Research Ethics Board does not encompass authorization to access the staff, students, facilities or resources of local institutions for the purposes of the research.

PLEASE NOTE you will need to submit the completed Delphi survey for review prior to data collection for that phase of the study.

Sincerely,

Stanley Varnhagen, PhD
Chair, Research Ethics Board 2 *Note: This correspondence includes an electronic signature (validation and approval via an online system).*

First Renewal

Date:	January 16, 2018
Amendment ID:	Pro00067856_REN1
Principal Investigator:	Peyman Azad Khaneghah
Study ID:	MS1_Pro00067856
Study Title:	A Rating Scale for Mental Health Mobile Applications for Older Adults
Supervisor:	Lili Liu
Approved Consent Form:	Approval Date Approved Document 2/9/2017 Phase 1 information letter and consent form 2/9/2017 phase 3 information letter and consent form 2/9/2017 Phase 2 information letter and consent form
Approval Expiry Date:	January 15, 2019

Thank you for submitting this renewal application. Your application has been reviewed and approved.

This re-approval is valid for one year. If your study continues past the expiration date as noted above, you will be required to complete another renewal request. Beginning at 30 days prior to the expiration date, you will receive notices that the study is about to expire. If you do not renew on or before the renewal expiry date, you will have to re-submit an ethics application.

Sincerely,

Stanley Varnhagen, PhD
Chair, Research Ethics Board 2

Note: This correspondence includes an electronic signature (validation and approval via an online system).

Second Renewal

Date:	January 2, 2019
Amendment ID:	Pro00067856_REN2
Principal Investigator:	Peyman Azad Khaneghah
Study ID:	MS4_Pro00067856
Study Title:	A Rating Index for Mobile Health Applications
Supervisor:	Lili Liu

Approved Consent Form:

Approval Date	Approved Document
11/28/2018	Consent_Phase3.3_23_11_2018
2/9/2017	Phase 1 information letter and consent form
8/9/2018	New document: information letter for paper based survey
2/9/2017	phase 3 information letter and consent form
2/9/2017	Phase 2 information letter and consent form
8/9/2018	Edited document: Previously approved letter clean copy
11/28/2018	Consent_Phase3.2_23_11_2018
11/28/2018	Consent_Phase3.1_23_11_2018

Approval Expiry Date: January 1, 2020

Thank you for submitting this renewal application. Your application has been reviewed and approved.

This re-approval is valid for one year. If your study continues past the expiration date as noted above, you will be required to complete another renewal request. Beginning at 30 days prior to the expiration date, you will receive notices that the study is about to expire. If you do not renew on or before the renewal expiry date, you will have to re-submit an ethics application.

Sincerely,

Ubaka Ogbogu, LLB, BL, LLM, SJD
Chair, Research Ethics Board 2

Note: This correspondence includes an electronic signature (validation and approval via an online system).

Appendix B: Information Letters and Consent Forms

Study Title: A Rating Scale for Mental Health Mobile Applications for Older Adults (Phase 1)

Research Investigator: Peyman Azad Khanegah
3-48, Corbett Hall
Faculty of Rehabilitation medicine
University of Alberta
Edmonton, AB, T6G 2G4
azadkhan@ualberta.ca
(780) 695-0447

Supervisor: Dr. Lili Liu
Department of Occupational Therapy
University of Alberta
Edmonton, AB, T6G 2G4
lili.liu@ualberta.ca
(780) 492-1728

Background

There are numerous mental health mobile applications (apps) available to the public. People can download these apps from online app markets. Few mental health apps are designed to take into consideration age-related cognitive, perceptual and sensory changes. We know very little about the quality of mental health apps. Currently, there are no scales to help users rate the quality of mental health apps.

Several stakeholders may benefit from using a scale to rate the quality of mental health apps. Examples are adults with a mental health condition, older adults, caregivers, clinicians, and app developers. We invite you to participate in this study because you belong to a stakeholder group that can use the scale. We want to know your inputs and suggestions about what is a good quality mental health mobile app.

Purpose

To create a scale that mental health service providers can use to rate the quality of mental health mobile apps for older adults.

Study Procedures

We ask you to participate in two focus groups.

In the first focus group, we will give you a tablet with three mental health apps installed on it. You will use each app for 5 minutes. After that, we will ask you to suggest items that should be on a scale to measure the quality of a mental health app. An example is "*how easy the app is to use"*. You may think of as many items as you can. We will audio record all of the suggestions and then transcribe them.

The second focus group will take place approximately two months after the first focus group. In the second focus group, we will ask you to categorize the items that you and other stakeholders suggested in the first focus group. We will write each item on a flash card. We will also write the name of each category on a large sheet of paper and place them in front of you. We will ask you to place each item under the category that best relates to it. For example, the item: "the app is easy to use" goes under the category "ease of use".

Each focus group will last for maximum of two hours. There will be between 4 to 6 participants in each focus group. In addition, at least two researchers will be present in each focus group. The focus groups will take place in Faculty of Rehabilitation Medicine at University of Alberta.

You can take a break at any time that you feel you need to. In addition, we will provide refreshments. We will audio record the focus groups and transcribe them. We will not collect any identifying information such as your name. If we find any identifying information such as names in the transcripts, we will remove them before conducting any analyses. We will review the transcripts and list all items that you think should be in an app quality rating scale.

Benefits

You may not directly benefit from participating in this study.

However, the use of mobile technology in healthcare provision is new. This project will add to this body of knowledge and will help us to develop a scale that clinicians can use to rate the quality of mental health mobile apps.

There will be no costs for your participation. We will provide the tablet with the applications already installed on them. We will also compensate parking or transportation costs.

Risk

To our knowledge, there are no risks in using the mobile devices or tablets. Many people on daily basis are using these devices.

Voluntary Participation

Your participation is voluntary.

You do not have to answer questions or participate if you feel uncomfortable and you can stop taking part at any time during the focus group.

You can withdraw from the study prior to the focus group, during the focus group, or up to two weeks after the focus groups. If you withdraw, we will attempt to remove your audio-recorded comments, but depending on the voice quality, this may not be possible.

After two weeks, we will have transcribed the audio recordings and removed all the identifying information from the transcripts. If you decide to withdraw from the study after this time, it is not possible for us to identify your comments to remove them from our results.

Confidentiality & Anonymity

The investigator will use the results of this study to write a PhD dissertation. We plan to present the findings in a scientific conference and use them to write a paper for publication. We will not use any information that may result in your identification in our presentations or publications.

The research team will make every effort to keep your information private. We will keep the information collected confidential, unless we are requested by law to reveal the information. As much as possible the information we keep will be anonymous; it will not have your name on it. We will guard your privacy as much as possible and use your information only for this project. We cannot guarantee that others in the focus group will maintain the confidentiality of what is said. Any electronic information will be stored at secure University of Alberta locations. The information will be password protected. Only the members of the research team and the Research Ethics Board at the University of Alberta will have access to this data. All records will be destroyed after five years. We may use the data we get from this study in future research, but if we do this, it will have to be approved by a Research Ethics Board.

Further Information

If you have any questions regarding this study, please contact:

Peyman Azad-Khaneghah, PhD candidate, Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 695-0447

Dr. John Misiaszek, Associate Dean Research, Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 492-2412

The plan for this study has been reviewed for its adherence to ethical guidelines by a Research Ethics Board at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Research Ethics Office at (780) 492-2615.

Consent Statement

I have read this form and the research study has been explained to me. I have been given the opportunity to ask questions and my questions have been answered. If I have additional questions, I have been told whom to contact. I agree to participate in the research study described above and will receive a copy of this consent form. I will receive a copy of this consent form after I sign it.

_____ _____

Participant's Name (printed) and Signature Date

_____ _____

Name (printed) and Signature of Person Obtaining Consent Date

Paper based Survey Information Letter (phase2)

Research Investigator: Peyman Azad Khanegah
3-48, Corbett Hall
Faculty of Rehabilitation medicine
University of Alberta
Edmonton, AB, T6G 2G4
azadkhan@ualberta.ca
(780) 695-0447

Supervisor: Dr. Lili Liu
Department of Occupational Therapy
University of Alberta
Edmonton, AB, T6G 2G4
lili.liu@ualberta.ca
(780) 492-1728

Background

The purpose of this study is to make a scale to help users rate the quality of wellness mobile applications (apps). There are many wellness apps that people can download from online app markets. We know very little about the quality of wellness apps. Right now, there are no scales to help users rate the quality of wellness apps.

You may have been involved in the first part of this study. In the first part we defined 10 quality indicators for wellness apps. Then we created a list of statements to rate apps for each indicator. Now, we want to evaluate the relevance of the statements to the quality indicators.

Study Procedures

We will use a Delphi technique for this study. We will use two or three rounds of surveys. The aim is to achieve consensus between participants about the relevance of the statements to the quality indicators. Completing the first survey will imply that you consent to participate. It also means that you are willing to receive the second survey. Each survey will have simple and specific instructions. We may give you a third and final survey if we do not achieve consensus after the second survey.

The amount of time necessary to complete this survey will vary with each participant, but it should not be more than 15 minutes. There will be no right or wrong answers. You only need to tell us how relevant you think each statement is to the given quality indicator.

Benefits

You may not directly benefit from participating in this study. The use of mobile apps in healthcare provision is new. This project will help us to make a scale for users to rate the quality of wellness mobile apps.

Risk

To our knowledge, there are no risks in participating in Delphi surveys. There will be no costs for your participation.

Voluntary Participation

It is voluntary to participate in this study. You do not have to participate if you feel uncomfortable.

You do not have to answer questions that make you uncomfortable. If you do not agree to participate you will not receive any more surveys from us.

If you decide to leave the study after you completed a survey, you have up to one week to inform us. To do so, contact Peyman Azad by e-mail (azadkhan@ualberta.ca) or phone (780-695 0447) and we will remove your survey from the study. After one week, we will remove all the identifying information from the surveys. At this point it will be impossible for us to find your completed survey and remove it from the study.

Confidentiality & Anonymity

The investigator will use the results of this study to write a PhD dissertation. Moreover, we may present the findings in a scientific conference and use them to write a peer reviewed

paper. We will not use any information that may result in your identification in our presentations or publications.

The research team will make every effort to keep your information private. We will use your responses only for this project. Any electronic information will be stored at secure University of Alberta locations. The information will be password protected. We will remove all identifying information from the responses before conducting any analysis. None of this electronic information will include your name or personal information. All records will be destroyed after five years. We may use the data we get from this study in future research, but if we do this, it will have to be approved by a Research Ethics Board.

Further Information

If you have any questions regarding this study, please contact:

- Peyman Azad-Khaneghah, AGE-WELL HQP and PhD candidate, Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 695-0447
- Dr. Lili Liu, Professor and Chair; Dept. of Occupational Therapy Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 492-1728
- Dr. John Misiaszek, Associate Dean Research, Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 492-2412

The plan for this study has been reviewed for its adherence to ethical guidelines by a Research Ethics Board at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Research Ethics Office at (780) 492-2615.

Consent Statement

I acknowledge that completing and submitting the survey implies that I have read the information letter for this study, I have been given the opportunity to ask questions, and my

questions have been answered. If I have additional questions, I have been told whom to contact.

I agree to participate in the research study described above.

Study Title: A Rating Index for Mobile Health Applications (Phase 3.1)

Research Investigator: Peyman Azad Khanegah
3-48, Corbett Hall
Faculty of Rehabilitation medicine
University of Alberta
Edmonton, AB, T6G 2G4
azadkhan@ualberta.ca
(780) 695-0447

Supervisors: Drs. Lili Liu
And Mary Roduta Roberts
Department of Occupational Therapy
University of Alberta
Edmonton, AB, T6G 2G4
lili.liu@ualberta.ca
Mary.Roberts@ualberta.ca

Background

The purpose of this study is to make an index to help end users such as older adults, patients, and clinicians rate the quality of mobile health applications (m-health apps). People can download many m-health apps from online app markets. We know very little about the quality of m-health apps. Right now, there are no Indices to help end users rate the quality of m-health apps.

In the previous phases of this study, we defined 10 constructs and 51 items to rate the quality of m-health apps. In this phase, we want to know if the number of items and the length of the index is acceptable to users of the index. Users include app users, technology developers, and health care providers.

We invite you to participate in this phase because we think you belong to one of the stakeholder groups who may use the index.

Purpose

To determine if the length of the index is acceptable to users of m-health apps including adults, technology developers, and healthcare providers.

Pre-focus group activity

We will ask you to complete an index review form before you participate in the focus group. The purpose of the review form is to prepare you for the focus group. You will receive the index review form at least one week before the focus group. We will ask you to complete this form and bring it with you to the focus group.

The objective of this exercise is to reduce the length of the index by removing unnecessary or redundant items from the index. On the index review form, you will see 10 quality domains for mobile health apps. Each quality domain has a number of items. We will ask you to review the items and tell us if we should keep or remove any of them from the index. Indicate your decision by marking "X" in the box provided on the review form. Note that while we want to reduce the length of the index, it must cover all 10-quality domains. Feel free to provide any comments in the comment section of the review form.

We will ask you to bring the index review forms with you to the focus group. The researcher will collect the form at the end of the focus group.

Focus group

There will be five or six other participants in the focus group, who like you, are stakeholders who would use the index. We will ask you to share with the rest of the participants in the focus group if you find the length of the index acceptable, and if not, which items you suggest we remove from the index. We will also ask you to provide a short rationale for your decision. Each participant will have 8 minutes for this task.

After everyone shares their choice and rational, the researcher will read each selected item and will ask participants to raise their hand if they agree that we should remove the item. The researcher will record the number of votes for each item.

After voting for all items, the group will take a break for 15 minutes. Meanwhile the researcher will remove the items that have received the majority of votes (at least 50% + 1) from the index and will print a copy of the new index.

After the break, the researcher will give each participant a copy of the new index to review. The group will spend 10 minutes for this task.

After the group reviewed the new index, the Researcher will ask participants to raise their hand if they do not find the new length of the index acceptable.

Next, the researcher will ask participants who do not find the new length acceptable to share which items they suggest to remove and their rational for this choice. The researcher will record the suggested items.

The researcher will read each item suggested and will ask the participants to raise their hand if they agree we should remove the item. The researcher will record the number of votes for each item and will remove items that receive the majority of votes (50%+1).

The focus group will be about two hours and will take place in Corbett Hall, Faculty of Rehabilitation Medicine-University of Alberta. The building is located on 8205, 114 Street, Edmonton, Alberta.

You can take a break at any time that you feel you need to. In addition, we will provide refreshments during the focus groups. We will audio record the focus group and transcribe it verbatim. The researchers will also take notes during the sessions. We will remove any identifying information such as your name before we analyze the notes or the transcripts.

Benefits

We will give a $20 gift card as a token of appreciation to you for participating in this focus group. We will also compensate for the parking costs at Corbett Hall and/or public transit to and from the study location. We will provide refreshments during the focus groups.

If you need to travel from another city to participate in the study, we will compensate for the round trip ground transportation and accommodation for up to 1 night.

Risk

To our knowledge, there are no risks in participating in focus groups.

Voluntary Participation

Your participation is voluntary.

You do not have to answer questions or participate if you feel uncomfortable and you can stop taking part at any time during the focus group.

You can withdraw from the study prior to the focus group, during the focus group, or up to two weeks after the focus groups. If you withdraw, we will attempt to remove your audio-recorded comments, but depending on the voice quality, this may not be possible.

After two weeks, we will have transcribed the audio recordings and removed all the identifying information from the transcripts. If you decide to withdraw from the study after this time, it is not possible for us to identify your comments to remove them from our results.

Confidentiality & Anonymity

The researcher will use the results of this study to write a PhD dissertation. We plan to present the findings in a scientific conference and use them to write a paper for publication. We will not use any information that may result in your identification in our presentations or publications.

The research team will make every effort to keep your information private. We will keep the information collected confidential, unless we are requested by law to reveal the information. As much as possible the information we keep will be anonymous; it will not have your name on it. We will guard your privacy as much as possible and use your information only for this project. We cannot guarantee that others in the focus group will maintain the confidentiality of what is said.

Any electronic information will be stored at secure University of Alberta locations. The information will be password protected. Only the members of the research team and the Research Ethics Board at the University of Alberta will have access to this data. All records will be destroyed after five years. We may use the data we get from this study in future research, but if we do this, it will have to be approved by a Research Ethics Board.

Further Information

If you have any questions regarding this study, please contact:

- Peyman Azad-Khaneghah, PhD candidate, Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 695-0447
- Dr. Lili Liu, Professor and Chair; Dept. of Occupational Therapy Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 492 0399
- Dr. Mary Roduta-Roberts, Assistant Professor; Dept. of Occupational Therapy Faculty of Rehabilitation Medicine, University of Alberta; Phone:(780) 492 0399
- Dr. John Misiaszek, Associate Dean Research, Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 492-2412

The plan for this study has been reviewed for its adherence to ethical guidelines by a Research Ethics Board at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Research Ethics Office at (780) 492-2615.

Consent Statement

I have read this form and the research study has been explained to me. I have been given the opportunity to ask questions and my questions have been answered. If I have additional questions, I have been told whom to contact. I agree to participate in the research study described above and will receive a copy of this signed consent form.

_____ _____

Participant's Name (printed) and Signature Date

_____ _____

Name (printed) and Signature of Person Obtaining Consent Date

Study Title: A Rating Index for Mobile Health Applications (Phase 3.2)

Research Investigator: Peyman Azad Khanegah
3-48, Corbett Hall
Faculty of Rehabilitation medicine
University of Alberta
Edmonton, AB, T6G 2G4
azadkhan@ualberta.ca
(780) 695-0447

Supervisors: Drs. Lili Liu
And Mary Roduta Roberts
Department of Occupational Therapy
University of Alberta
Edmonton, AB, T6G 2G4
lili.liu@ualberta.ca
Mary.Roberts@ualberta.ca

Background

The purpose of this study is to create an index for users such as older adults, patients, and health professionals to rate the quality of mobile health applications (m-health apps). People can download many m-health apps from online app markets. We know very little about the quality of m-health apps.

We invite you to participate in this study because we think you belong to one of the stakeholder groups who may use the index.

Purpose

To ensure that the items in this index used to rate the quality of m-health apps are clear and easy to understand.

Study Procedures

We ask you to participate in a 1-hour one-on-one interview to review the clarity of the items in the index that we have developed for quality and usability of m-health apps.

We ask you to bring your smart phone to the interview. We ask you to download and install an m-health app that is compatible with your smart phone. We will also provide you with a copy of the index to rate the quality and usability of the m-health app.

We ask you to tell us everything that comes to mind as you answer each item in the index. Particularly we are interested to know how you interpreted the item, if the item was easy to understand, if the scale responses were appropriate for the item, if you needed more prompts to answer the item, and if you recommend any revisions to improve the clarity of the item. The researcher may also provide prompts to help you if necessary.

The session will take place at the University of Alberta, Room 1-43 Corbett Hall, 8205, 114 Street, Edmonton, Alberta. The session will take 1 hour. You can take a break at any time.

The researcher will audio record the session and transcribe it verbatim. The researcher will also observe how you apply the index and will take notes. The researcher will remove any identifying information such as your name from the notes and transcripts.

Benefits

We will give a $20 gift card as a token of appreciation to you for participating in this study.

Risk

To our knowledge, there are no risks in using smart phones or completing the rating scale.

Voluntary Participation

Your participation is voluntary.

You do not have to answer questions or participate if you feel uncomfortable and you can stop taking part at any time during the session.

You can withdraw from the study prior to the interview session, during the interview, or up to two weeks after the interview. If you withdraw within two weeks, we will remove your audio-recorded comments. After two weeks, we will have completed revising the index. At this

point, since we have merged your data with data collected from other participants, it is not possible for us to reliably identify your comments and remove them from our results.

Confidentiality & Anonymity

The researcher will use the results of this study to write a PhD dissertation. We plan to present the findings in a scientific conference and use them to write a paper for publication. We will not use any information that may result in your identification in our presentations or publications.

The research team will make every effort to keep your information private. We will keep the information collected confidential, unless we are requested by law to reveal the information. As much as possible the information we keep will be anonymous; it will not have your name on it. We will guard your privacy as much as possible and use your information only for this project.

Any electronic information will be stored at secure University of Alberta locations. The information will be password protected. Only the members of the research team and the Research Ethics Board at the University of Alberta will have access to this data. All records will be destroyed after five years. We may use the data we get from this study in future research, but if we do this, we will seek approval from the Research Ethics Board.

Further Information

If you have any questions regarding this study, please contact:

- Peyman Azad-Khaneghah, PhD candidate, Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 695-0447
- Dr. Lili Liu, Professor and Chair; Dept. of Occupational Therapy Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 492 0399

- Dr. Mary Roduta-Roberts, Assistant Professor; Dept. of Occupational Therapy Faculty of Rehabilitation Medicine, University of Alberta; Phone:(780) 492 0399
- Dr. John Misiaszek, Associate Dean Research, Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 492-2412

The plan for this study has been reviewed for its adherence to ethical guidelines by a Research Ethics Board at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Research Ethics Office at (780) 492-2615.

Consent Statement

I have read this form and the research study has been explained to me. I have been given the opportunity to ask questions and my questions have been answered. If I have additional questions, I have been told whom to contact. I agree to participate in the research study described above and will receive a copy of this signed consent form.

_____ _____

Participant's Name (printed) and Signature Date

_____ _____

Name (printed) and Signature of Person Obtaining Consent Date

Study Title: A Rating Index for Mobile Health Applications (Phase 3.3)

Research Investigator: Peyman Azad Khanegah
3-48, Corbett Hall
Faculty of Rehabilitation medicine
University of Alberta
Edmonton, AB, T6G 2G4
azadkhan@ualberta.ca
(780) 695-0447

Supervisors: Drs. Lili Liu
And Mary Roduta Roberts
Department of Occupational Therapy
University of Alberta
Edmonton, AB, T6G 2G4
lili.liu@ualberta.ca
Mary.Roberts@ualberta.ca

Background

The purpose of this study is to create an index for users such as older adults, patients, and health professionals to rate the quality of mobile health applications (m-health apps). People can download many m-health apps from online app markets. We know very little about the quality of m-health apps.

We invite you to participate in this study because we think you belong to one of the stakeholder groups who may use the index.

Purpose

To investigate the reliability and validity of the index to rate the quality and usability of m-health apps.

Study Procedures

The researcher will meet with you in a one-on-one session to teach you how to use the indices for this study. You will need to bring your smart phone with you to this practice session. The researcher will ask you to download and install two m-health apps that are compatible with your smart phone. He will ask you to apply the index that we have developed to rate the quality and usability of the apps. The name of this scale is Alberta Index for Quality and Usability of m-Health Apps (the new index, for short). Next, the researcher will ask you to rate the same apps

once more, but using another index called Mobile Application Rating Scale for Users (U-MARS). The researcher will answer any questions you may have on how to use each of the indices.

After learning to use the two indices, the researcher will give you a list of 10 more m-health apps that are compatible to your smart phone. The researcher will also give you 10 paper copies of the new index and 10 copies of U-MARS.

You will take the list of the apps and the copies of the indices home. The apps on the list are numbered from 1 to 10. You will need to download each app on your smart phone. For apps numbered 1 to 5, first use the new index to rate the quality of the app and write the total score on the paper copy of the index. Next, use U-MARS to rate the quality of the same app and write the total score in the specified space on the index. For apps numbered 6 to 10, first use U-MARS to rate their quality and after that, use the new index to rate the quality of the same apps. Once again, make sure you write down the total scores in the specified spaces on the paper copies of the indices.

It will take a total of 30 minutes to rate each app and complete both indices. You will have 30 days to complete all ratings. During this time, if you have any questions or difficulties with using the indices you may contact the researcher by phone or e-mail. The researcher will also contact you (by phone or e-mail) every 7 days to follow up on your progress and answer any questions or concerns you may have. After 30 days, the researcher will meet with you and collect the completed indices.

Benefits

We will give a $30 gift card as a token of appreciation to you for participating in this study. We will also compensate the costs of apps that you need to purchase for this study.

Risk

To our knowledge, there are no risks in using smart phones or completing the rating scale.

Voluntary Participation

Your participation is voluntary.

You do not have to participate if you feel uncomfortable and you can stop taking part at any time during the study.

You can withdraw your participation during the study or up to one week after you gave the completed indices to the researcher. If you withdraw, we will remove your completed indices from the study.

After 1 week, we will have removed all the identifying information from the indices. If you decide to withdraw from the study after this time, it is not possible for us to identify which indices you completed to remove them from our results.

Confidentiality & Anonymity

The researcher will use the results of this study to write a PhD dissertation. We plan to present the findings in a scientific conference and use them to write a paper for publication. We will not use any information that may result in your identification in our presentations or publications.

The research team will make every effort to keep your information private. We will keep the information collected confidential, unless we are requested by law to reveal the information. As much as possible the information we keep will be anonymous; it will not have your name on it. We will guard your privacy as much as possible and use your information only for this project.

Any electronic information will be stored at secure University of Alberta locations. The information will be password protected. Only the members of the research team and the Research Ethics Board at the University of Alberta will have access to this data. All records will be destroyed after five years. We may use the data we get from this study in future research, but if we do this, it will have to be approved by a Research Ethics Board.

Further Information

If you have any questions regarding this study, please contact:

- Peyman Azad-Khaneghah, PhD candidate, Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 695-0447
- Dr. Lili Liu, Professor and Chair; Dept. of Occupational Therapy Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 492 0399
- Dr. Mary Roduta-Roberts, Assistant Professor; Dept. of Occupational Therapy Faculty of Rehabilitation Medicine, University of Alberta; Phone:(780) 492 0399
- Dr. John Misiaszek, Associate Dean Research, Faculty of Rehabilitation Medicine, University of Alberta; Phone: (780) 492-2412

The plan for this study has been reviewed for its adherence to ethical guidelines by a Research Ethics Board at the University of Alberta. For questions regarding participant rights and ethical conduct of research, contact the Research Ethics Office at (780) 492-2615.

Consent Statement

I have read this form and the research study has been explained to me. I have been given the opportunity to ask questions and my questions have been answered. If I have additional questions, I have been told whom to contact. I agree to participate in the research study described above and will receive a copy of this signed consent form.

_____	_____
Participant's Name (printed) and Signature	Date
_____	_____
Name (printed) and Signature of Person Obtaining Consent	Date

Appendix C: Paper based survey

Section 1. Demographics

First Name: ---------- Last Name: ----------

E-mail address: ----------

Age range: 20-29 30-39 40-49 50-59 60 and above

Which of the following describes you:
1) I am an occupational Therapist
2) I am a psychologist
3) I am a psychiatrist
4) I am a nurse
5) I am an app developer
6) I am an app user
7) Others (Please specify) ----------

How many years have you used a smart phone? ---------- Years

How many years have you used a tablet? ----------Years

Section 2.

Instructions. Please rate the relevance of each statement to the given quality indicator. The options are 1) Not relevant 2) Relevant 3) Unable to assess relevance.

1. **Privacy**

For each statement, rate its relevance for evaluating whether the app protects the privacy of its users (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
1. The app has a privacy policy that indicates what information is collected by the app, who will have access to this information, and how this information will be used.			
2. The app description indicates compliance with a health information act or privacy regulation (e.g., Health Information Act of Canada)			
3. The app asks users to enter personal information such as name, date of birth, home address, or phone number.			
4. The app asks users to enter personal financial information such as credit card number or debit card number.			
5. The app requests access to phone sensors such as the microphone, camera, or location of the phone.			
6. The app collects users' password for social media (e.g., for Face book, Twitter, Instagram) or e-mail.			

Suggest revisions to any of the statements in this list. (Use back of this page if you need more space)
Suggest statements that are missing in this list. (Use back of this page if you need more space)

2. Security

For each statement, rate its relevance for evaluating whether the app uses ways to secure users' information (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
7. The app uses various ways to protect users' information that is recorded on the smart phone (e.g., log in password)			
8. The app protects users' information that is recorded on the app website or cloud.			
9. The app protects users' information that is transmitted (e.g., via the internet, Wi-Fi, or Bluetooth)			
10. The app notifies users if another person accesses their account from a different device			

11. The app requires users to create a password before starting to use the app.			
12. The app requires users to create a password before starting to record their data on a cloud or a website.			
13. The app description states that the app encrypts users' information that is recorded or transmitted.			
Suggest revisions to any of the statements in this list. (Use back of this page if you need more space)			
Suggest statements that are missing in this list. (Use back of this page if you need more space)			

3. Trustworthiness

For each statement, rate its relevance for evaluating whether an app provides up-to-date, credible, and trustworthy information (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
14. The app has been updated less than a year ago.			
15. The app was developed by credible developer(s) based on the information on the app description (e.g., government health services).			
16. The app description mentions health care professionals were involved in development of the app.			
17. The app description mentions potential users (e.g., older adults, patients, etc.) were involved in development of the app.			
18. The app provides evidence on its effectiveness (e.g., link to scientific studies).			
19. The app provides evidence of approval from a third party such as Health Canada or FDA.			
20. The app clearly states the source of the health information that it provides.			

21. The app clearly specifies its financial sponsors.			
22. The paid version of the app advertises a particular health care product (e.g., medication, assistive device, meal plan, food supplement)			

3. Trustworthiness (continued)

For each statement, rate its relevance for evaluating whether a health app provides up-to-date, credible, and trustworthy information (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
23. The app content can cause harm (e.g., create suicidal ideas, encourage drug abuse)			
24. The app discloses any potential risks that users may face if they use the content of the app.			
25. There is evidence of conflict of interest (e.g., a pharmaceutical company has developed a depression app that advertises a certain anti-depressant medication)			
26. Conflict of interest, if present, are declared.			

Suggest revisions to any of the statements in this list. (Use back of this page if you need more space)
Suggest statements that are missing in this list. (Use back of this page if you need more space)

4. Ease of Use

For each statement, rate the relevance of the statement to evaluating whether the app is easy to use (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
27. In general, the app is easy to use.			

28. The app provides information that is easy to understand.			
29. The app uses icons that are easy to understand.			
30. It is easy to navigate different pages of the app.			
31. The app uses font size that is easy to see			
32. The app uses colours that are easy to see.			
33. The app uses icons that are easy to see.			
34. The app's touchable elements are large enough to be touched using one finger.			
35. The app provides a link for technical support to the users.			
36. Mistakes are easy to fix in the app (e.g., entering wrong information or deleting data)			
37. The app lets me know if I enter wrong information.			

38. The app confirms with user before information is removed or deleted.			
39. In general, it is easy to make changes to the information that a user enters in the app.			
40. The app is accessible for those who have visual impairments.			
41. The app is accessible for those who have hearing impairments.			
Suggest revisions to any of the statements in this list. (Use back of this page if you need more space)			
Suggest statements that are missing in this list. (Use back of this page if you need more space)			

5. User engagement

For each statement, rate its relevance for evaluating whether an app is engaging and/or enjoyable to use (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
42. The app features are interactive.			
43. The app features are engaging.			
44. The app uses various ways such as notifications, reminders, or achievement points to engage users.			
45. The app content is presented in an interesting way (e.g., contains the right mix of video/audio/text/graphics).			
46. The app attracts users to use it regularly			
47. The app provides feedback on user's actions (e.g., providing prompts, notifications, and reminders).			

Suggest revisions to any of the statements in this list. (Use back of this page if you need more space)
Suggest statements that are missing in this list. (Use back of this page if you need more space)

6. Aesthetics

For each statement, rate its relevance for evaluating aesthetics in an app. (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
48. The app uses colours that are attractive.			
49. Layout design: the arrangement of elements (buttons/icons/menus/content) on the screen are Attractive.			
50. Visual Design: the visual presentation of the app is attractive.			

51. Graphics: The quality/resolution of graphics used for buttons/icons/menus/content is high.			
52. In general, the app's appearance is appropriate for the target users.			
Suggest revisions to any of the statements in this list. (Use back of this page if you need more space)			
Suggest statements that are missing in this list. (Use back of this page if you need more space)			

7. Customizability

For each statement, rate its relevance for evaluating whether features of an app are customizable (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
53. User can change features of the app (e.g., font type, font Size, font colour, background colour, size of the images, icons, colour of icons, notifications, sound/music, Language).			
54. User can zoom in or out to see the content better.			
55. User can change the app language to the one that she/ he prefers.			
56. The app retains all necessary settings for features of the app (e.g. sound, content, notifications, etc.).			
57. The app helps automate some tasks, (e.g., pre-populated fields, suggestions based on previous inputs)			
Suggest revisions to any of the statements in this list. (Use back of this page if you need more space)			

Suggest statements that are missing in this list. (Use back of this page if you need more space)

8. Functionality

For each statement, rate the relevance for evaluating the functionality of an app (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
58. The app components work correctly (e.g., Does not crash, all links work)			
59. The app needs internet connection to work.			
60. The app sends information to other phones, apps, or devices.			
61. The app receives information from other phones, apps, or devices (e.g., Fitbit or smart watches).			
62. The app is available for different operating systems (e.g., android and Apple).			
63. The app is available for different devices (e.g., phone, tablet, desktop computer).			

Suggest revisions to any of the statements in this list. (Use back of this page if you need more space)
Suggest statements that are missing in this list. (Use back of this page if you need more space)

9. Affordability

For each statement, rate its relevance for evaluating whether an app is affordable (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
64. The app has a free trial version.			
65. The app is free			
66. A recurrent (e.g., monthly, annually) membership fee is required to keep using the app.			
67. The app is affordable.			

Suggest revisions to any of the statements in this list. (Use back of this page if you need more space)
Suggest statements that are missing in this list. (Use back of this page if you need more space)

10. Usefulness

For each statement, rate its relevance for evaluating whether an app is useful (Put X in the box provided)	**Not relevant**	**Relevant**	**Unable to assess relevance**
68. The purpose of the app is clearly stated or implied in its name or description in the app description.			
69. The app content includes what is implied in the name or described in the app description.			
70. The app provides useful health information.			
71. The app is useful for self-monitoring symptoms (e.g., blood sugar, blood pressure, mood, anxiety levels, pain)			

72. The app is useful as a reminder (e.g., for medications/appointments).			
73. The app is useful for making appointments with health care professionals.			
74. The app is useful to communicate with health care professionals.			
Suggest revisions to any of the statements in this list. (Use back of this page if you need more space)			
Suggest statements that are missing in this list. (Use back of this page if you need more space)			

Appendix D: Index Review Form

ID…….

This review form is designed to help you prepare for the focus group on …….

Please bring this completed from to the focus group.

The objective of this exercise is to reduce the length of the index or the number of items. Please review each item and suggest whether it should be kept or removed by marking "X" in the box. Note that the index should cover all quality domains from "Privacy" to "Usefulness". Feel free to add your comments in the provided space.

A) Privacy: Items in this construct measure how transparent m-health apps are about the user information that the app collects.	**Keep**	**Remove**
1. The app has a privacy policy		
2. The app's privacy policy explains what information is collected by the app		
3. The app's privacy policy explains who will have access to the information collected by the app		
4. The app's privacy policy explains how the information collected by the app will be used		
5. The app requests access to phone sensors such as the microphone, camera, or location of the phone.		
Comments:		

B) Security: Items in this construct look at the measures m-health apps take to prevent unauthorized access to app users' information.	**Keep**	**Remove**
6. The app uses various ways to protect users' information that is recorded on the smart phone (e.g., login password).		
7. The app uses various ways to protect users' information that it records on the app website or cloud (e.g., log in password or data encryption).		
8. The app notifies users if another person accesses their account from a different device.		
9. The app protects users' information that is transmitted (e.g., via the internet, Wi-Fi, or Bluetooth).		
10. The app description states that the app encrypts users' information that is recorded or transmitted.		
11. The app requires users to create a password before starting to use the app.		
Comments:		
C) Trustworthiness: Items in this construct measure if m-health apps provide any evidence that the information they provide are reliable and trustworthy.	**Keep**	**Remove**
12. The app description mentions that health care professionals were involved in development of the app's content.		
13. The app clearly states the source of the health information that it provides.		

Trustworthiness (continued): Items in this construct measure if m-health apps provide any evidence that the information they provide are reliable and trustworthy.	**Keep**	**Remove**
14. The app declares any conflict of interest, if present.		
15. The app was developed by credible developer(s) based on the information on the app description (e.g., government health services).		
16. The app provides evidence of approval from a credible third party (examples are governmental organizations, not for profit organizations, hospitals, or universities).		
17. There is evidence of conflict of interest in the app (e.g., a pharmaceutical company has developed a depression app that advertises a certain anti-depressant medication).		
18. The app content can cause harm (e.g., create suicidal ideas, encourage drug abuse).		
19. The app discloses any potential risks that users may face if they use the content of the app.		
Comments:		

D) Ease of use: Items in this construct measure if users perceive m-health apps easy to use.	**Keep**	**Remove**
20. It is easy to navigate different pages of the app.		
21. It is easy to make changes to the information that a user enters in the app.		
22. The app uses icons that are easy to understand.		
23. The app uses icons that are easy to see.		
24. It is easy to fix mistakes in the app (e.g., entering wrong information or deleting data).		
25. In general, the app is easy to use.		
26. The app provides information that is easy to understand.		
27. The app uses default font size that is easy to see.		
28. The app's touchable elements are large enough to be touched using one finger.		
29. The app confirms with user before information is removed or deleted.		
30. The app uses colours that are easy to see.		
Comments:		

E) User engagement: Items in this construct measure the degree to which m-health apps engage users.	**Keep**	**Remove**
31. The app content is presented in an interesting way (e.g., contains the right mix of video, audio, text and graphics).		
32. In general the app features are engaging.		
33. The app uses various ways such as notifications, reminders, or achievement points to engage users.		
Comments:		
F) Aesthetics: Items in this construct measure if m-health apps are visually appealing. Elements considered are colours used, size and type of fonts, size and shape of the icons used, and arrangement of the elements	**Keep**	**Remove**
34. The quality of the graphics used in the app is high.		
35. In general, the app's appearance is appropriate for the target users.		
36. The visual presentation of the app is attractive.		
37. The arrangement of elements such as buttons, icons, menus, or content on the screen are Attractive.		
Comments:		

G) Customizability: Items in this construct measure the extent a user can modify settings of m-health apps.	**Keep**	**Remove**
38. User can zoom in or out to see the content better.		
39. The app retains all necessary settings made by the user for features of the app (e.g. sound, notifications, etc.).		
40. User can change features of the app (e.g., font type, font Size, font colour, background colour, size of the images, icons, colour of icons, notifications, sound/music, Language).		
41. User can change the app's language to the one that she/he prefers.		
42. The app helps automate some tasks, (e.g., pre-populated fields, suggestions based on previous inputs).		
Comments:		
H) Functionality: Items in this construct measure if m-health apps work properly.	**Keep**	**Remove**
43. The app is available for different operating systems (e.g., android and Apple).		
44. The app is available for different devices (e.g., phone, tablet, desktop computer).		
45. The app components work correctly (e.g., Does not crash, all links work).		

Functionality (Continued): Items in this construct measure if m-health apps work properly.	**Keep**	**Remove**
46. The app receives information from other phones, apps, or devices (e.g., Fitbit or smart watches).		
Comments:		
I) Affordability: Items in this construct measure if users perceive the costs of using m-health apps is worth the benefits.	**Keep**	**Remove**
47. The app is affordable.		
48. The app has a free version.		
49. A recurrent (e.g., monthly, annually) membership fee is required to keep using the app.		
Comments:		

J) Usefulness: Items in this construct measure the degree to which users perceive m-health apps useful.	**Keep**	**Remove**
50. The app provides useful health information.		
51. The app is useful for self-monitoring symptoms (e.g., blood sugar, blood pressure, mood, anxiety levels, or pain).		
Usefulness (Continued): Items in this construct measure the degree to which users perceive m-health apps useful.	**Keep**	**Remove**
52. The app content includes what is implied in the name or described in the app description.		
53. The app is useful as a reminder (e.g., for medications or appointments).		
Comments:		

Appendix E: Pilot Draft of Alberta Rating Index for Apps (ARIA-Pilot)

App Name: ..... I have used this app before: Yes ☐ No ☐

Use this index to rate quality of mobile health apps that are accessible on app stores. The index has two parts.

Part A: Complete this part before downloading the app on your phone or tablet. Read the app description on the app store and rate the extent to which the following statements are true about the app.	Not at all true	Slightly true	Somewhat true	Mostly true	Completely true	Not Applicable
1. The description provided for the app fits my goals. (You may find this information under **"About this app"** on Google Play; or **"Preview"** on iTunes. Read about the **goals** and **objectives** of the app, **price**, and **in-app purchases**)	0	1	2	3	4	NA
2. Based on the description provided on the app or the app developer's website, I can tell that the app has been developed by relevant experts in the field. (To go to the developers' website, tap **"Visit webpage" on Google play**; and tap **"Developer Website" on iTunes**.)	0	1	2	3	4	NA
3. The app description includes a statement on risks associated with using the app.	0	1	2	3	4	NA
4. The app declares any conflict of interest, if present.	0	1	2	3	4	NA
5. The app has a privacy policy that explains (1) what information is collected by the app, (2) who will have access to this information, (3) how this information will be used. (Look for the "**Privacy policy**" of the app under the app description on app stores. On **Google play** it is shown with an icon that looks like a padlock "🔒". On **iTunes** the icon looks like a hand "✋")	0	1	2	3	4	NA
6. Considering the costs associated with using the app, including in-app purchases and subscription renewal fees, I find the app affordable to use.	0	1	2	3	4	NA
Part A Score: ...../ 24						

Now download the app and complete part B of the index on the back of this page.

Part B: Use the app for as long as you feel is necessary to become familiar with its features (at least 10 minutes). Try all **the links** and **buttons** on the screen. Pay attention to **icons**, **colors**, **font size**, **the information** provided in the app, and **looks and feels** of the app. If applicable, go to the **"Settings"** of the app and see if the app allows you to **customize** the settings (such as font size, reminders, and notifications). After you became familiar with the app, rate the extent to which the following statements are true about the app. **You may go back to the app and check the features if you need**.	Not at all true	Slightly true	Somewhat true	Mostly true	Completely true	Not Applicable
7. The app uses at least one security measure such as user name and password or biometric identifiers (finger prints, face recognition) to allow users access the app.	0	1	2	3	4	NA
8. The app asks my consent to access the phone's camera, microphone, my location, or my documents (contacts, photos).	0	1	2	3	4	NA
9. The app clearly provides the source of all health information that it provides.	0	1	2	3	4	NA
10. It is easy to move from one page to another.	0	1	2	3	4	NA
11. It is easy to see components of the app such as texts, icons, and buttons.	0	1	2	3	4	NA
12. It is easy to understand the information provided by the app (in the form of text, graphs, tables, audio, or video).	0	1	2	3	4	NA
13. The app components work correctly with no errors (e.g., does not crash, all links work).	0	1	2	3	4	NA
14. I can customize the app settings to my satisfaction.	0	1	2	3	4	NA
15. The content of the app is appropriate for the intended users of the app.	0	1	2	3	4	NA
16. The app can help me to achieve my goals.	0	1	2	3	4	NA
17. It is pleasing to use the app.	0	1	2	3	4	NA
18. I am satisfied with using the app.	0	1	2	3	4	NA
Part B Score: ...../48						

Appendix F: Alberta Rating Index for Apps-Users version

App Name: ..... I have used this app before: Yes ☐ No ☐

Use this index to rate the quality of a mobile health app. Please complete Part A and Part B.

	Part A: **Complete this part before downloading the app on your phone or tablet.** • First, **find the app on your online app store**. For Apple products open "App store" for Android products go to the "Play Store." • Then, read the app description from the online app store and rate the extent to which you agree with the following statements.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Purpose	1. The description of the app's purpose fits my goals. **(You may find app description under "About this app" on Google Play; or "Preview" on iTunes)**	0	1	2	3	4
Trustworthiness	2. Based on the description provided on the app store or the app developer's website, I trust that relevant experts in the field have developed the app.	0	1	2	3	4
	3. The app description includes a statement about the risks associated with using the app.	0	1	2	3	4
	4. The app declares conflicts of interest, if any.	0	1	2	3	4
Privacy	5. The app has a privacy policy that explains: (1) what information is collected by the app, (2) who will have access to this information, and (3) how this information will be used. **(Look for the "Privacy policy" of the app under the app description on "App Store" for Apple products or "Play store" for Android products)**	0	1	2	3	4
Affordability	6. The costs associated with using the app, including in-app purchases and subscription renewal fees, are affordable. **(Go to the "App Store" for Apple products or "Play store" for Android products to learn how much does it cost to use the app)**	0	1	2	3	4
	Add up the scores for Part A:...../24					

	Part B: Use the app for at least 10 minutes or as long as you feel it is necessary to become familiar with its features. Try all **the links** and **buttons** on the screen. After you become familiar with the app, rate the extent to which you agree with the following statements. **You may go back to the app and check the features**.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Security	1. The app uses at least one security measure, such as **user name and password** or biometric identifiers **(fingerprints, face recognition)**, to allow users to access the app.	0	1	2	3	4
	2. The app asks for my consent if it needs to access the phone's camera, microphone, my location, my contacts, or my photos.	0	1	2	3	4
Trustworthiness	3. The app mentions the references for the health information that it provides. Examples of trustworthy references are scientific papers or websites of the governmental health organizations, universities, or not-for-profit health groups.	0	1	2	3	4
Ease of use	4. Moving from one screen of the app to another is easy for me.	0	1	2	3	4
	5. It is easy for me to see components of the app such as text, icons, and buttons. **Pay attention to colours and sizes.**	0	1	2	3	4
	6. It is easy for me to understand the information provided by the app. **Pay attention to the text, graphs, tables, audio, or video.**	0	1	2	3	4
Functionality	7. The app components work correctly. For example, it does not crash or all links work.	0	1	2	3	4
	8. I can customize the app settings to my satisfaction **(Try to customize language, font size, font colour, background colour, reminders, and notifications).**	0	1	2	3	4
Target users	9. The content of the app is appropriate for me considering my age, gender, education, and cultural background.	0	1	2	3	4
Usefulness and Satisfaction	10. The app can help me to achieve my goals.	0	1	2	3	4
	11. The app is pleasing to use.	0	1	2	3	4
	12. I am satisfied with using the app.	0	1	2	3	4
	Add up the scores for Part B:...../48					
	Circle the number of stars that best represents your overall rating for quality of this app: **(1 star =Worst app I have ever used; 5 Stars= Best app I have ever used)**	★	★	★	★	★

Indice d'évaluation de la qualité des applications de santé mobiles (version pour les utilisateurs)

Nom de l'application ..... J'ai déjà utilisé cette application : Oui □ Non □

Utilisez ce tableau pour évaluer la qualité d'une application de santé mobile. Merci de remplir les Parties A et B.

Partie A: **Remplir cette partie avant de télécharger l'application sur votre téléphone ou votre tablette.** • Tout d'abord, **trouvez l'application en ligne**. Pour les produits Apple, ouvrez « App store »; pour les produits Android, rendez-vous au « Play Store » • Puis, lisez la description de l'application et indiquez dans quelle mesure vous êtes d'accord (ou non) avec les énoncés suivants.		Fortement en désaccord	Pas d' accord	Neutre	D' accord	Fortement d' accord
Objectif	1. La description de l'objectif de l'application répond à mes objectifs. **(vous trouverez la description de l'application sous « About this app » (Au sujet de cette app) dans Google Play; ou sous « Preview » (Aperçu) dans iTunes)**	0	1	2	3	4
Fiabilité	2. Selon la description fournie dans l'App Store ou sur le site Web du développeur, j'ai confiance que l'application a été développée par des experts compétents dans le domaine.	0	1	2	3	4
	3. La description de l'application comprend un énoncé portant sur les risques associés à l'utilisation de l'application.	0	1	2	3	4
	4. Le cas échéant, les conflits d'intérêts sont énoncés.	0	1	2	3	4
Confidentialité	5. La politique de confidentialité de l'application énonce : (1) les renseignements recueillis par l'application; (2) qui aura accès à ces informations; et (3) comment ces informations seront utilisées. (**recherchez la politique de confidentialité dans la description de l'application dans l'App Store pour les produits Apple ou dans le Play store pour les produits Android)**	0	1	2	3	4
Abordabilité	6. Les coûts associés à l'utilisation de l'application, y compris les achats intégrés et les frais de renouvellement d'adhésion, sont abordables. **(rendez-vous dans l'App Store pour les produits Apple ou dans le Play store pour les produits Android pour savoir combien coûte l'utilisation de l'application)**	0	1	2	3	4
Additionnez les scores de la Partie A:...../24						

Partie B : Utilisez l'application pendant au moins 10 minutes, ou pendant aussi longtemps que vous l'estimez nécessaire pour vous familiariser à ses caractéristiques. Essayez tous **les liens** et **les boutons** sur l'écran. Après vous être familiarisé avec l'application, évaluez la mesure dans laquelle vous êtes d'accord ou non avec les énoncés suivants. **Vous pouvez retourner à l'application et vérifier ses caractéristiques**.		Fortement en désaccord	Pas d' accord	Neutre	D' accord	Fortement d' accord
Sécurité	1. L'application est munie d'au moins une mesure de sécurité, comme **un nom d'utilisateur et un mot de passe**, ou encore des identificateurs biométriques **(empreintes digitales, reconnaissance faciale)** qui permettent aux utilisateurs d'accéder à l'application.	0	1	2	3	4
	2. L'application demande mon consentement si elle doit accéder à ma caméra, mon micro, mon emplacement, mes contacts ou mes photos.	0	1	2	3	4
Fiabilité	3. L'application indique les références des informations relatives à la santé qui y sont mentionnées. Les exemples de références fiables sont les articles scientifiques ou les sites Web des organismes de santé gouvernementaux, les universités ou les groupes sans but lucratif œuvrant dans le domaine de la santé.	0	1	2	3	4
Facilité d' utilisation	4. Je trouve qu'il est facile de passer d'un écran à l'autre de l'application.	0	1	2	3	4
	5. Je trouve qu'il est facile de voir les éléments de l'application, comme le texte, les icônes et les boutons. **Faites attention aux couleurs et aux tailles.**	0	1	2	3	4
	6. Je trouve qu'il est facile de comprendre les informations fournies par l'application. **Faites attention au texte, aux graphiques, aux tableaux, à l'audio ou à la vidéo.**	0	1	2	3	4
Fonctionnalité	7. Les éléments de l'application fonctionnent correctement. Par exemple, le système ne plante pas et tous les liens fonctionnent correctement.	0	1	2	3	4
	8. Je peux personnaliser les paramètres de l'application **(essayez de personnaliser la langue, la taille et la couleur de la police, la couleur de l'arrière-plan, les rappels et les notifications)** à mon goût**).**	0	1	2	3	4
Utilisateurs ciblés	9. Le contenu de l'application est approprié pour moi, compte tenu de mon âge, mon sexe, ma formation et de mes antécédents culturels.	0	1	2	3	4
Utilité et satisfaction	10. L'application peut m'aider à atteindre mes objectifs.	0	1	2	3	4
	11. L'application est agréable à utiliser.	0	1	2	3	4
	12. Je suis satisfait de l'utilisation de l'application.	0	1	2	3	4
Additionnez les scores de la Partie B:...../48						
Entourez le nombre d'étoiles qui représente au mieux votre évaluation générale de la qualité de l'application: **(1 étoile = La pire application que j'ai jamais utilisée; 5 étoiles = La meilleure application que j'ai jamais utilisée)**		★ ★ ★ ★ ★				

Appendix G: Alberta Rating Index for Apps-Care providers

App Name: ..... I have used this app before: Yes ☐ No ☐

Use this index to rate the quality of a mobile health app for use by a client or patient (referred to as user). Please complete Part A and Part B.

	Part A: **Complete this part before downloading the app on your phone or tablet.** • First, **find the app on your online app store**. For Apple products open "App store" for Android products go to the "Play Store." • Then, read the app description from the online app store and rate the extent to which you agree with the following statements.	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Purpose	1. The description of the app's purpose fits the user's goals. **(You may find app description under "About this app" on Google Play; or "Preview" on iTunes)**	0	1	2	3	4
Trustworthiness	2. Based on the description provided on the app store or the app developer's website, the user can trust that relevant experts in the field have developed the app.	0	1	2	3	4
	3. The app description includes a statement about the risks associated with using the app.	0	1	2	3	4
	4. The app declares conflicts of interest, if any.	0	1	2	3	4
Privacy	5. The app has a privacy policy that explains: (1) what information is collected by the app, (2) who will have access to this information, and (3) how this information will be used. **(Look for the "Privacy policy" of the app under the app description on "App Store" for Apple products or "Play store" for Android products)**	0	1	2	3	4
Affordability	6. The costs associated with using the app, including in-app purchases and subscription renewal fees, are affordable for the user. **(Go to the "App Store" for Apple products or "Play store" for Android products to learn how much does it cost to use the app)**	0	1	2	3	4
Add up the scores for Part A:...../24						

Part B: Use the app for at least 10 minutes or as long as you feel it is necessary to become familiar with its features. Try all **the links** and **buttons** on the screen. After you become familiar with the app, rate the extent to which you agree with the following statements. **You may go back to the app and check the features**.		Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Security	1. The app uses at least one security measure, such as **user name and password** or biometric identifiers **(fingerprints, face recognition)**, to allow the user to access the app.	0	1	2	3	4
	2. The app asks for the user's consent if it needs to access the phone's camera, microphone, user's location, contacts, or photos.	0	1	2	3	4
Trustworthiness	3. The app mentions the references for the health information that it provides. Examples of trustworthy references are scientific papers or websites of the governmental health organizations, universities, or not-for-profit health groups.	0	1	2	3	4
Ease of use	4. Moving from one screen of the app to another would be easy for the user.	0	1	2	3	4
	5. It would be easy for the user to see components of the app such as text, icons, and buttons. **Pay attention to colours and sizes.**	0	1	2	3	4
	6. It would be easy for the user to understand the information provided by the app. **Pay attention to the text, graphs, tables, audio, or video.**	0	1	2	3	4
Functionality	7. The app components work correctly. For example, it does not crash or all links work.	0	1	2	3	4
	8. The user would be able to customize the app settings **(language, font size, font colour, background colour, reminders, and notifications)** to her/his satisfaction.	0	1	2	3	4
Target users	9. The content of the app is appropriate for the user considering her/his age, gender, education, and cultural background.	0	1	2	3	4
Usefulness and Satisfaction	10. The app can help the user to achieve her/his goals.	0	1	2	3	4
	11. The user would find the app pleasing to use.	0	1	2	3	4
	12. The user would be satisfied with using the app.	0	1	2	3	4
Add up the scores for Part B: ……../48						
Overall, I would recommend using this app to the user.		0	1	2	3	4
Circle the number of stars that best represents your overall rating for quality of this app: **(1 star =Worst app I have ever used; 5 Stars= Best app I have ever used)**		★	★	★	★	★

Indice d'évaluation de la qualité des applications de santé mobiles (version pour les fournisseurs de soins)

Nom de l'application : ..... J'ai déjà utilisé cette application : Oui ☐ Non ☐

Utilisez ce tableau pour évaluer la qualité d'une application de santé mobile chez un client ou un patient (ci-après, « utilisateur »). Merci de remplir les Parties A et B.

Partie A: **Remplir cette partie avant de télécharger l'application sur votre téléphone ou votre tablette.** • Tout d'abord, **trouvez l'application en ligne**. Pour les produits Apple, ouvrez « App store »; pour les produits Android, rendez-vous au « Play Store ». • Puis, lisez la description de l'application et indiquez dans quelle mesure vous êtes d'accord (ou non) avec les énoncés suivants.		Fortement en désaccord	Pas d' accord	Neutre	D' accord	Fortement d' accord
Objectif	1. La description de l'objectif de l'application convient aux objectifs de l'utilisateur. **(vous trouverez la description de l'application sous « About this app » *(Au sujet de cette app)* dans Google Play; ou sous « Preview » *(Aperçu)* dans iTunes)**	0	1	2	3	4
Fiabilité	2. Selon la description fournie dans l'App Store ou sur le site Web du développeur, l'utilisateur peut avoir confiance que l'application a été développée par des experts compétents dans le domaine.	0	1	2	3	4
	3. La description de l'application comprend un énoncé portant sur les risques associés à l'utilisation de l'application.	0	1	2	3	4
	4. Le cas échéant, les conflits d'intérêts sont énoncés.	0	1	2	3	4
Confidentialité	5. La politique de confidentialité de l'application énonce : (1) les renseignements recueillis par l'application; (2) qui aura accès à ces informations; et (3) comment ces informations seront utilisées. **(recherchez la politique de confidentialité dans la description de l'application dans l'App Store pour les produits Apple ou dans le Play store pour les produits Android)**	0	1	2	3	4
Abordabilité	6. Les coûts associés à l'utilisation de l'application, y compris les achats intégrés et les frais de renouvellement d'adhésion, sont abordables pour l'utilisateur. **(rendez-vous dans l'App Store pour les produits Apple ou dans le Play store pour les produits Android pour savoir combien coûte l'utilisation de l'application)**	0	1	2	3	4
Additionnez les scores de la Partie A:...../24						

Partie B : Utilisez l'application pendant au moins 10 minutes, ou pendant aussi longtemps que vous l'estimez nécessaire pour vous familiariser à ses caractéristiques. Essayez tous **les liens** et **les boutons** sur l'écran. Après vous être familiarisé avec l'application, évaluez la mesure dans laquelle vous êtes d'accord ou non avec les énoncés suivants. **Vous pouvez retourner à l'application et vérifier ses caractéristiques**.		Fortement en désaccord	Pas d' accord	Neutre	D' accord	Fortement d' accord
Sécurité	1. L'application est munie d'au moins une mesure de sécurité, comme un **nom d'utilisateur et un mot de passe,** ou encore des identificateurs biométriques **(empreintes digitales, reconnaissance faciale)** qui permettent à l'utilisateur d'accéder à l'application.	0	1	2	3	4
	2. L'application demande le consentement de l'utilisateur si elle doit accéder à sa caméra, son micro, son emplacement, ses contacts ou ses photos.	0	1	2	3	4
Fiabilité	3. L'application indique les références des informations relatives à la santé qui y sont mentionnées. Les exemples de références fiables sont les articles scientifiques ou les sites Web des organismes de santé gouvernementaux, les universités ou les groupes sans but lucratif œuvrant dans le domaine de la santé.	0	1	2	3	4
Facilité d' utilisation	4. L'utilisateur passera facilement d'un écran à l'autre de l'application.	0	1	2	3	4
	5. L'utilisateur verra facilement les éléments de l'application, comme le texte, les icônes et les boutons. **Faites attention aux couleurs et aux tailles.**	0	1	2	3	4
	6. L'utilisateur comprendra facilement les informations fournies par l'application. **Faites attention au texte, aux graphiques, aux tableaux, à l'audio ou à la vidéo.**	0	1	2	3	4
Fonctionnalité	7. Les éléments de l'application fonctionnent correctement. Par exemple, le système ne plante pas et tous les liens fonctionnent correctement.	0	1	2	3	4
	8. L'utilisateur pourra personnaliser les paramètres de l'application **(langue, taille et couleur de la police, couleur de l'arrière-plan, rappels et notifications)** à son goût.	0	1	2	3	4
Utilisateurs ciblés	9. Le contenu de l'application est approprié pour l'utilisateur, compte tenu de son âge, son sexe, sa formation et de ses antécédents culturels.	0	1	2	3	4
Utilité et satisfaction	10. L'application peut aider l'utilisateur à atteindre ses objectifs.	0	1	2	3	4
	11. L'utilisateur trouvera l'application agréable à utiliser.	0	1	2	3	4
	12. L'utilisateur sera satisfait d'utiliser l'application.	0	1	2	3	4
Additionnez les scores de la Partie B : ……./48						
En général, je recommanderais cette application à l'utilisateur.		0	1	2	3	4
Entourez le nombre d'étoiles qui représente au mieux votre évaluation générale de la qualité de l'application : **(1 étoile = La pire application que j'ai jamais utilisée; 5 étoiles = La meilleure application que j'ai jamais utilisée)**		★	★	★	★	★

Appendix H: Mobile Application Rating Scale: User version (u-MARS)

Instructions for use:

Raters should:

1. Use the app and trial it thoroughly for at least 10 minutes;

2. Determine how easy it is to use, how well it functions and does it do what it purports to do;

3. Review app settings, developer information, external links, security features, etc.

Scoring

A: Engagement Mean Score = _____

B: Functionality Mean Score = _____

C: Aesthetics Mean Score = _____

D: Information Mean Score* = _____

* Exclude questions rated as "N/A" from the mean score calculation.

App quality mean score _____ = A + B + C + D / 4

App Name: _____

Circle the number that most accurately represents the quality of the app you are rating. All items are rated on a 5-point scale from "1.Inadequate" to "5.Excellent". Select N/A if the app component is irrelevant.

App Quality Ratings

SECTION A

Engagement – fun, interesting, customisable, interactive, has prompts (e.g. sends alerts, messages, reminders, feedback, enables sharing)

1. Entertainment: Is the app fun/entertaining to use? Does it have components that make it more fun than other similar apps?

1. Dull, not fun or entertaining at all
2. Mostly boring
3. OK, fun enough to entertain user for a brief time (< 5 minutes)
4. Moderately fun and entertaining, would entertain user for some time (5-10 minutes total)
5. Highly entertaining and fun, would stimulate repeat use

2. Interest: Is the app interesting to use? Does it present its information in an interesting way compared to other similar apps?

1. Not interesting at all
2. Mostly uninteresting
3. OK, neither interesting nor uninteresting; would engage user for a brief time (< 5 minutes)
4. Moderately interesting; would engage user for some time (5-10 minutes total)
5. Very interesting, would engage user in repeat use

3. Customisation: Does it allow you to customise the settings and preferences that you would like to (e.g. sound, content and notifications)?

1. Does not allow any customisation or requires setting to be input every time
2. Allows little customisation and that limits app's functions
3. Basic customisation to function adequately
4. Allows numerous options for customisation
5. Allows complete tailoring the user's characteristics/preferences, remembers all settings

4. Interactivity: Does it allow user input, provide feedback, contain prompts (reminders, sharing options, notifications, etc.)?

1. No interactive features and/or no response to user input
2. Some, but not enough interactive features which limits app's functions
3. Basic interactive features to function adequately
4. Offers a variety of interactive features, feedback and user input options
5. Very high level of responsiveness through interactive features, feedback and user input options

5. Target group: Is the app content (visuals, language, design) appropriate for the target audience?

1. Completely inappropriate, unclear or confusing
2. Mostly inappropriate, unclear or confusing
3. Acceptable but not specifically designed for the target audience. May be inappropriate/ unclear/confusing at times
4. Designed for the target audience, with minor issues
5. Designed specifically for the target audience, no issues found

SECTION B

Functionality – app functioning, easy to learn, navigation, flow logic, and gestural design of app

6. Performance: How accurately/fast do the app features (functions) and components (buttons/menus) work?

1. App is broken; no/insufficient/inaccurate response (e.g. crashes/bugs/broken features, etc.)
2. Some functions work, but lagging or contains major technical problems
3. App works overall. Some technical problems need fixing, or is slow at times
4. Mostly functional with minor/negligible problems
5. Perfect/timely response; no technical bugs found, or contains a 'loading time left' indicator (if relevant)

7. Ease of use: How easy is it to learn how to use the app; how clear are the menu labels, icons and instructions?

1. No/limited instructions; menu labels, icons are confusing; complicated
2. Takes a lot of time or effort
3. Takes some time or effort
4. Easy to learn (or has clear instructions)
5. Able to use app immediately; intuitive; simple (no instructions needed)

8. Navigation: Does moving between screens make sense; Does app have all necessary links between screens?

1. No logical connection between screens at all /navigation is difficult
2. Understandable after a lot of time/effort
3. Understandable after some time/effort

4. Easy to understand/navigate
5. Perfectly logical, easy, clear and intuitive screen flow throughout, and/or has shortcuts

9. Gestural design: Do taps/swipes/pinches/scrolls make sense? Are they consistent across all components/screens?

1. Completely inconsistent/confusing
2. Often inconsistent/confusing
3. OK with some inconsistencies/confusing elements
4. Mostly consistent/intuitive with negligible problems
5. Perfectly consistent and intuitive

SECTION C

Aesthetics – graphic design, overall visual appeal, colour scheme, and stylistic consistency

10. Layout: Is arrangement and size of buttons, icons, menus and content on the screen appropriate?

1. Very bad design, cluttered, some options impossible to select, locate, see or read
2. Bad design, random, unclear, some options difficult to select/locate/see/read
3. Satisfactory, few problems with selecting/locating/seeing/reading items
4. Mostly clear, able to select/locate/see/read items
5. Professional, simple, clear, orderly, logically organised

11. Graphics: How high is the quality/resolution of graphics used for buttons, icons, menus and content?

    1. Graphics appear amateur, very poor visual design - disproportionate, stylistically inconsistent
    2. Low quality/low resolution graphics; low quality visual design – disproportionate
    3. Moderate quality graphics and visual design (generally consistent in style)
    4. High quality/resolution graphics and visual design – mostly proportionate, consistent in style
    5. Very high quality/resolution graphics and visual design - proportionate, consistent in style throughout

12. Visual appeal: How good does the app look?

    1. Ugly, unpleasant to look at, poorly designed, clashing, mismatched colours
    2. Bad – poorly designed, bad use of colour, visually boring
    3. OK – average, neither pleasant, nor unpleasant
    4. Pleasant – seamless graphics – consistent and professionally designed
    5. Beautiful – very attractive, memorable, stands out; use of colour enhances app features/menus

SECTION D

Information – Contains high quality information (e.g. text, feedback, measures, references) from a credible source.

13. Quality of information: Is app content correct, well written, and relevant to the goal/topic of the app?

N/A There is no information within the app

1 Irrelevant/inappropriate/incoherent/incorrect

2 Poor. Barely relevant/appropriate/coherent/may be incorrect

3 Moderately relevant/appropriate/coherent/and appears correct

4 Relevant/appropriate/coherent/correct

5 Highly relevant, appropriate, coherent, and correct

14. Quantity of information: Is the information within the app comprehensive but concise?

N/A There is no information within the app

1. Minimal or overwhelming
2. Insufficient or possibly overwhelming
3. OK but not comprehensive or concise
4. Offers a broad range of information, has some gaps or unnecessary detail; or has no links to more information and resources
5. Comprehensive and concise; contains links to more information and resources

15. Visual information: Is visual explanation of concepts – through charts/graphs/images/videos, etc. – clear, logical, correct?

N/A There is no visual information within the app (e.g. it only contains audio, or text)

1. Completely unclear/confusing/wrong or necessary but missing
2. Mostly unclear/confusing/wrong
3. OK but often unclear/confusing/wrong
4. Mostly clear/logical/correct with negligible issues
5. Perfectly clear/logical/correct

16. Credibility of source: does the information within the app seem to come from a credible source?

N/A There is no information within the app

1. Suspicious source
2. Lacks credibility
3. Not suspicious but legitimacy of source is unclear
4. Possibly comes from a legitimate source
5. Definitely comes from a legitimate/specialised source

App subjective quality

SECTION E

17. Would you recommend this app to people who might benefit from it?

1. Not at all. I would not recommend this app to anyone
2. There are very few people I would recommend this app to
3. Maybe. There are several people I would recommend this app to
4. There are many people I would recommend this app to
5. Definitely. I would recommend this app to everyone

18. How many times do you think you would use this app in the next 12 months if it was relevant to you?

1. None
2. 1-2
3. 3-10
4. 10-50
5. 5 >50

19. Would you pay for this app?

1 Definitely not

2

3

4

5 Definitely yes

20. What is your overall (star) rating of the app?

⋆ One of the worst apps I've used

⋆⋆

⋆⋆⋆ Average

⋆⋆⋆⋆

⋆⋆⋆⋆⋆ One of the best apps I've used

Appendix I: Mock Patient Profile

The following is a profile for a hypothetical patient. Consider this information when you rate the quality of the apps in this study.

- Gender: Male Age: 65 years
- Height: 170 Cm (5' 6") Weight: 78 Kg (172 lb)
- Language(s) spoken: English Annual income: $20,000 (AISH)
- Education: High school diploma
- Occupation: Retired
- Marital status: Single (Never Married)
- Living arrangement: Lives alone
- Housing: Lives in a subsidized condo
- **Medical History**
- Diagnosis: Bipolar mood disorder (Was diagnosed 40 years ago)
- Chief complaint(s): Poor sleep, Forgets to take medication on time, tired, and low energy during the day.
- The patient is not a smoker, has no known substance abuse problems, and is not suffering from any eating disorders.
- Medications: Mood stabilizers and Antidepressant-antipsychotic medication.
- Psychotherapy: Cognitive behavioral therapy (CBT)
- Current Mental Status
- Suicidal ideation: NO
- Delusions/Hallucinations: NO
- Insight: Has insight into his condition, Motivated to get better

Appendix J: Items that did not fit any criteria

1. Detail of depth [of wound] meant for ID physician [151]
2. Filling in a mEMA app questionnaire was an interruption of my daily activities [127]
3. After the researcher's explanation I understood how the app would work [127]
4. Filling in 1 mEMA app questionnaire took too long [127]
5. I carried my smartphone with me every day [127]
6. I filled in the mEMA app questionnaires 4 times a day [127]
7. I filled in the mEMA app questionnaires for 7 consecutive days [127]
8. It is easy to carry the smartphone with me [127]
9. Fast achievement of a first feeling of success [128]
10. It is easy to access the app [133]
11. The app is an effective strategy for discharge education [133]
12. Overall impression [130]
13. Does it provide a mechanism to turn off app advertising [144]
14. Does it show in the download what content or services offered by the app (App Store) [144]
15. Does the app check the size of the download [144]
16. Does the app indicate the user profile to which it is addressed? (professional/patient) [144]
17. Does the app Web site belong to a partner [144]
18. In the app, is there information on the accessibility features [144]
19. Is the level of evidence or grade of recommendation used indicated [144]

20. Is there an age verification process for the user [144]
21. Felt confident [44]
22. use if prescribed [44]
23. well organized [44]
24. Does this app provide an explanation for its grading system either in the app or on their website [148]
25. Does this app use an evidence grading system [148]
26. How many topics does this app indicate it contains on their website [148]
27. What is the method of evidence collection the app uses to create the information in their topics (Look for the method on the website) [148]
28. Advertisement [61]
29. Application comprehensiveness [61]
30. Other components (images, videos, special features) [61]
31. Text Search Fields [61]
32. Do long tasks show non-stationary activity indicators [123]
33. Is the login delayed to allow the user to use a particular functionality first [123]
34. Would you change or add something from this application [123]
35. Is there any specific clinical scenario or patient subgroup in which using the app seems particularly likely to be useful [69]
36. I would use this app if it were available at my health facility [59]
37. Does app categorize interactions [between medications] according to severity [63]
38. Does app include information for special populations (e.g., pediatrics, pregnancy, renal/liver dysfunction) [63]

39. Adaption of existing data collection systems [117]
40. Adequate sample sizes for evaluation ensured [117]
41. Are novel BCTs employed by app to be evaluated [117]
42. Collected data available to relevant bodies [117]
43. Consider times when users more open to change [117]
44. Detailed description of the app publicly available [117]
45. Does the evaluation use qualitative tools [117]
46. Establishment of routine data collection [117]
47. Evaluation planned prior to release of app [117]
48. Is a description of the evaluation design available [117]
49. Is it clear which BCT was employed [117]
50. Process and outcomes using objective measures [117]
51. Publicly available manual [117]
52. Specialist input for evaluation [117]
53. Take account of users' environment [117]
54. Uptake and reach of the app assessed [117]
55. Use of one or more recognized BCT [117]
56. Will intervention fidelity be evaluated [117]
57. Will outcomes be assessed for over 1 year [117]
58. Will the efficacy of the app be evaluated [117]
59. Will the impact on health inequalities be assessed [117]
60. Definition of enuresis in accordance with ICCS (2010) [International Children's Continence Society (ICCS)] [142]

61. I would prefer to use a mobile app-based reporting system, such as the AEFI app, instead of a web-based system on a computer [135]
62. For programs explicitly designed to be used by minors, the system includes a section requiring the approval/supervision of a legal guardian [64]
63. Relatability. Does the EHP offer a good representation of a human factor that is easily relatable within the therapeutic context/process? Examples include a professional character who directs the user throughout the program; a peer who was in a similar situation and is now better (e.g., fitness);

    a vivid virtual character who leads the user; a community of people working together for change. [64]
64. Does app contain a disclaimer or similar statement, clearly declaring that the app is not meant to replace the advice of a health professional?

    (0) No disclaimer stated

    (1) Yes, a disclaimer / similar statement exists [66]
65. Does the app have snooze/postpone function?

    (0) No, it does not have snooze and postpone functions

    (1) Yes, it has snooze/postpone function [66]
66. Will you do something differently after using this app [65]
67. Will you try to do something new after using this app [65]
68. Topics covered in enough detail [65]
69. Is there a statement within the app about the developer's commitment to addressing problems reported to them? (eg, timescales to respond, commitment to eradicate reported bugs and faults) [70]

Appendix K: Content validation index (CVI) for items

Item number	Criteria	Item	CVI
1	Privacy	The app has a privacy policy that indicates what information is collected by the app, who will have access to this information, and how this information will be used [43, 63, 64, 66, 68-70, 120, 140, 144].	1
2	Privacy	The app description indicates compliance with a health information act or privacy regulation (e.g., Health Information Act of Canada) [64, 68].	0.72
3	Privacy	The app asks users to enter personal information such as name, date of birth, home address, or phone number [43, 69].	0.61
4	Privacy	The app asks users to enter personal financial information such as credit card number or debit card number [43, 69].	0.72
5	Privacy	The app requests access to phone sensors such as the microphone, camera, or location of the phone.	0.78
6	Privacy	The app collects users' password for social media (e.g., for Face book, Twitter, Instagram) or e-mail [64].	0.61
7	Security	The app uses various ways to protect users' information that is recorded on the smart phone (e.g., log in password) [43, 63, 64, 67, 68, 70, 122, 144, 146, 156]	0.94
8	Security	The app protects users' information that is recorded on the app website or cloud [144].	0.94
9	Security	The app protects users' information that is transmitted (e.g., via the internet, Wi-Fi, or Bluetooth) [64, 70, 144].	0.89
10	Security	The app notifies users if another person accesses their account from a different device [64, 144, 156].	0.94
11	Security	The app requires users to create a password before starting to use the app [63, 64, 67, 144].	0.83
12	Security	The app requires users to create a password before starting to record their data on a cloud or a website [64].	0.94
13	Security	The app description states that the app encrypts users' information that is recorded or transmitted [156].	0.89

Item number	Criteria	Item	CVI
14	Trustworthiness	The app has been updated less than a year ago [63, 64, 117, 122, 144, 146, 148, 157].	0.67
15	Trustworthiness	The app was developed by credible developer(s) based on the information on the app description (e.g., government health services) [5, 63, 64, 66, 67, 69, 85, 120, 144, 148, 157].	0.83
16	Trustworthiness	The app description mentions health care professionals were involved in development of the app [63, 64, 66, 68, 85, 117, 144, 148, 157].	0.89
17	Trustworthiness	The app description mentions potential users (e.g., older adults, patients, etc.) were involved in development of the app [64, 66, 70, 85, 117, 144, 148, 157].	0.72
18	Trustworthiness	The app provides evidence on its effectiveness (e.g., link to scientific studies) [5, 63, 64, 69, 70, 85, 117, 148].	0.78
19	Trustworthiness	The app provides evidence of approval from a third party such as Health Canada or FDA [63, 64, 69, 70, 85, 117].	0.83
20	Trustworthiness	The app clearly states the source of the health information that it provides [61, 63, 64, 67-69, 120, 144, 146, 148, 157].	0.89
21	Trustworthiness	The app clearly specifies its financial sponsors [63, 66, 120, 144, 157].	0.72
22	Trustworthiness	The paid version of the app advertises a particular health care product (e.g., medication, assistive device, meal plan, food supplement) [120, 144]	0.50
23	Trustworthiness	There is evidence of conflict of interest (e.g., a pharmaceutical company has developed a depression app that advertises a certain anti-depressant medication) [63, 66, 120, 157]	0.83
24	Trustworthiness	Conflict of interest, if present, are declared [63, 66, 120, 157].	0.89
25	Trustworthiness	The app content can cause harm (e.g., create suicidal ideas, encourage drug abuse) [117]	0.83
26	Trustworthiness	The app discloses any potential risks that users may face if they use the content of the app [117, 144].	0.83
27	Ease of Use	In general, the app is easy to use [5, 44, 55, 57, 59, 63, 64, 66-68, 85, 111, 127, 128, 131, 137, 141, 142, 146, 149, 151, 158-161].	0.89

Item number	Criteria	Item	CVI
28	Ease of Use	The app provides information that is easy to understand [44, 57, 59, 63, 64, 67, 68, 70, 85, 111, 122, 128, 130, 131, 133, 140-142, 159, 160].	0.89
29	Ease of Use	The app uses icons that are easy to understand [67, 68, 122, 128, 140, 141, 146].	0.94
30	Ease of Use	It is easy to navigate different pages of the app [5, 63-65, 67, 85, 122, 130, 131, 133, 143, 146, 158, 161].	1
31	Ease of Use	The app uses font size that is easy to see [63, 64, 67, 68, 133]	0.89
32	Ease of Use	The app uses colours that are easy to see [146].	0.83
33	Ease of Use	The app uses icons that are easy to see.	0.94
34	Ease of Use	The app's touchable elements are large enough to be touched using one finger [64].	0.89
35	Ease of Use	The app provides a link for technical support to the users [63, 69, 70, 122, 144, 146, 151, 162].	0.72
36	Ease of Use	Mistakes are easy to fix in the app (e.g., entering wrong information or deleting data) [44, 57, 59, 128, 131, 159]	0.94
37	Ease of Use	The app lets me know if I enter wrong information [122, 131].	0.78
38	Ease of Use	The app confirms with user before information is removed or deleted [123].	0.89
39	Ease of Use	In general, it is easy to make changes to the information that a user enters in the app [44, 128].	1
40	Ease of Use	The app is accessible for those who have visual impairments [70, 162].	0.67
41	Ease of Use	The app is accessible for those who have hearing impairments [70, 162].	0.61
42	User engagement	The app features are interactive [5, 64].	0.78
43	User engagement	The app features are engaging [142].	0.83

Item number	Criteria	Item	CVI
44	User engagement	The app uses various ways such as notifications, reminders, or achievement points to engage users [5, 64, 67].	0.72
45	User engagement	The app content is presented in an interesting way (e.g., contains the right mix of video/audio/text/graphics) [64, 127].	0.89
46	User engagement	The app attracts users to use it regularly [44, 133, 160, 163].	0.56
47	User engagement	The app provides feedback on user's actions (e.g., providing prompts, notifications, and reminders) [65, 67, 122, 158].	0.72
48	Aesthetics	The app uses colours that are attractive [68, 123].	0.67
49	Aesthetics	Layout design: the arrangement of elements (buttons/icons/menus/content) on the screen are Attractive [44, 59, 63, 64, 68, 69, 122, 128, 133, 146, 158, 164].	0.78
50	Aesthetics	Visual Design: the visual presentation of the app is attractive [5, 64, 122, 133, 137, 146, 158].	0.83
51	Aesthetics	Graphics: The quality/resolution of graphics used for buttons/icons/menus/content is high [5, 85].	0.94
52	Aesthetics	In general, the app's appearance is appropriate for the target users [5, 63, 69, 85, 131, 144, 146, 151, 162].	0.94
53	Customizability	User can change features of the app (e.g., font type, font Size, font colour, background colour, size of the images, icons, colour of icons, notifications, sound/music, Language) [63, 64, 67, 128, 146].	0.83
54	Customizability	User can zoom in or out to see the content better [5].	0.89
55	Customizability	User can change the app language to the one that she/he prefer [142, 146].	0.83
56	Customizability	The app retains all necessary settings for features of the app (e.g. sound, content, notifications, etc.) [5].	0.89
57	Customizability	The app helps automate some tasks, (e.g., pre-populated fields, suggestions based on previous inputs) [64, 66, 67, 117, 122, 144, 146]	0.83
58	Functionality	The app components work correctly (e.g., Does not crash, all links work) [5, 44, 63, 65, 67, 85, 122, 127, 144]	0.94

Item number	Criteria	Item	CVI
59	Functionality	The app needs internet connection to work [61, 63, 67].	0.78
60	Functionality	The app sends information to other phones, apps, or devices [61, 63, 67, 142, 146].	0.61
61	Functionality	The app receives information from other phones, apps, or devices (e.g., Fitbit or smart watches) [61, 63, 67].	0.83
62	Functionality	The app is available for different operating systems (e.g., android and Apple) [144].	1.00
63	Functionality	The app is available for different devices (e.g., phone, tablet, desktop computer) [144].	1.00
64	Affordability	The app has a free trial version [61].	0.78
65	Affordability	The app is free [61].	0.83
66	Affordability	A recurrent (e.g., monthly, annually) membership fee is required to keep using the app [61].	0.78
67	Affordability	The app is affordable [61].	1.00
68	Usefulness	The purpose of the app is clearly stated or implied in its name or description in the app description [5, 54, 64-66, 69, 70, 117, 120, 133, 144, 146].	0.78
69	Usefulness	The app content includes what is implied in the name or described in the app description [5, 65, 131, 146].	0.83
70	Usefulness	The app provides useful health information [44, 65, 66, 130, 132, 133].	1.00
71	Usefulness	The app is useful for self-monitoring symptoms (e.g., blood sugar, blood pressure, mood, anxiety levels, pain) [57, 63, 66, 67, 117, 122, 132]	0.94
72	Usefulness	The app is useful as a reminder (e.g., for medications/appointments) [66, 67].	0.83
73	Usefulness	The app is useful for making appointments with health care professionals [66, 117].	0.72
74	Usefulness	The app is useful to communicate with health care professionals [66, 111, 120, 123].	0.67