Effectiveness of a chatbot in improving the mental wellbeing of health workers in Malawi during the COVID-19 pandemic: A randomized, controlled trial

Short title: Chatbot use and mental wellbeing of health workers

Eckhard F. Kleinau[1*], Tilinao Lamba[2], Wanda Jaskiewicz[3], Katy Gorentz[3], Ines Hungerbuehler[4], Donya Rahimi[3], Demoubly Kokota[2], Limbika Maliwichi[2], Edister S. Jamu[2], Alex Zumazuma[5], Mariana Negrão[4], Raphael Mota[4], Yasmine Khouri[4], Michael Kapps[4]

[1] Global-Latin America and Caribbean (LAC)-West Africa Region, University Research Co. (URC), Chevy Chase, Maryland, United States of America

[2] Dept. of Psychology, University of Malawi – Chancellor College, Zomba, Malawi

[3] Global Health Division, Chemonics International, Washington, District of Columbia, United States of America

[4] Clinical Division, Vitalk, São Paulo, Brazil

[5] Department of Mental Health, Kamuzu University of Health Sciences (KUHES), Blantyre, Malawi

* Corresponding author

E-mail: ekleinau@urc-chs.com (EK)

Abstract

After the deleterious effects of the COVID-19 pandemic on healthcare worker mental health, we tested the effectiveness of an interactive chatbot, Vitalk, for improving wellbeing and resilience among healthcare workers in Malawi, a country with few mental health professionals. We conducted a randomized, controlled trial (RCT) to investigate our hypothesis that Vitalk is more effective in improving mental health and resilience outcomes than passive Internet resources.

For our 2-arm, 8-week, parallel RCT (ISRCTN Registry: trial ID ISRCTN16378480), we recruited participants from 8 professional cadres from public and private healthcare facilities. The treatment arm used Vitalk; the control arm received links to Internet resources. Of 1,584 participants, 512 completed baseline and endline assessments. Six assessments provided outcome measures for: anxiety (GAD-7); depression (PHQ-9); burnout (OLBI); loneliness (ULCA); resilience (RS-14); and resilience-building activities. We analyzed effectiveness using mixed-effects linear models, effect size estimates, and reliable change in risk levels.

Results from mixed-model analyses support our hypothesis. Difference-in-differences estimators showed that Vitalk reduced: depression (-0.68 [95% CI -1.15 to -0.21]); anxiety (-0.44 [95% CI -0.88 to 0.01]); and burnout (-0.58 [95% CI -1.32 to 0.15]). Changes in resilience (1.47 [95% CI 0.05 to 2.88]) and resilience-building activities (1.22 [95% CI 0.56 to 1.87]) were significantly greater in the treatment group. We observed no treatment effect on loneliness. Our RCT produced a medium effect size.

This is the first RCT of a mental health app for healthcare workers during the pandemic in Southern Africa combining multiple mental wellbeing outcomes, and measuring resilience and resilience-building activities. A significant number of participants could have benefited from mental health support (1 in 8

43 reported anxiety and depression; 3 in 4 suffered burnout; and 1 in 4 had low resilience). Such help is not

44 readily available in Malawi. Vitalk has the potential to fill this gap.

45

Introduction

Healthcare provision can be stressful even under normal circumstances, and maintaining the mental wellbeing of health workers is of utmost importance for optimal and safe patient care [1]. The COVID-19 pandemic has overwhelmed countries' health systems and increased care-related pressure to ensure patient and staff safety. Health workers confront life and death decisions, physical exhaustion, lack of protective equipment, and fear of infection as daily threats to their mental wellbeing. An increasing number of studies report a high proportion of health professionals globally suffering from depression, anxiety, and burnout. A systematic review of 59 studies found that a median of 24% of health workers suffered from anxiety, 21% from depression, and 37% from distress [2]. A meta-analysis from 36 countries reported slightly higher levels (8.0% for depression; 26.9% for anxiety; 24.1% for post-traumatic stress symptoms; 36.5% for stress; and 50.0% for distress) [3]. Recent studies from sub-Saharan Africa showed similar levels. A multi-center cross-sectional study from Ghana, found the following levels: depression (21.1%), anxiety (27.8%) and stress (8.2%) [4]. Research from Ethiopia found the following prevalence: depression (20.2%), anxiety (21.9%), and psychological distress (15.5%) [5]. Data from South Africa suggest high levels of mental disorders (around 50%) among all types of health workers due to work-related stress [6]. A study from Malawi using a small sample of nurses and the Coronavirus Anxiety Scale suggests that 26% (n=26) of respondents had COVID-19-related anxiety and 48% (n=49) functional impairment [7].

Mental disorders not only pose threats to patient safety and health workers' quality of life, but they also incur high economic costs. A World Health Organization (WHO)-led study, well before the COVID-19 pandemic, found that depression and anxiety disorders cost the global economy US$1 trillion each year [8]. This study estimated a return on investments of between 2.3 and 5.7 to 1 if treatment of these mental disorders was scaled up. Mental health interventions include basic psychosocial counseling for

69 milder cases, and more intensive psychosocial treatment plus antidepressant drugs for more severe
70 cases. Over the last decade, computer- or Internet-based cognitive behavioral therapy (c-CBT or i-CBT)
71 has been tested and implemented as alternatives to in-person treatment. Research has shown that c-
72 CBT-based self-administered interventions improve depression and anxiety in adults. A meta-analysis of
73 49 randomized controlled trials (RCT) revealed a significant medium to large effect size (Hedges' g=0.77,
74 95% CI 0.59 to 0.95) of c-CBT for depression and anxiety [9].

75 Another meta-analysis of 22 RCTs found an even greater effect size (g=0.88, 95% CI 0.76 to 0.99) [10].
76 However, a recent systematic review found small to medium post-treatment pooled effect sizes
77 regarding depressive symptoms (g=0.51, 95% CI 0.30 to 0.72) and anxiety symptoms (g=0.44, 95% CI
78 0.23 to0.65) of c-CBT for reducing these symptoms in adolescents and young adults compared to passive
79 controls [11]. Clinical trials have established that mobile applications can effectively deliver CBT
80 programs for the treatment of: depression [12]; self-management of chronic pain conditions [13]; and
81 social anxiety disorder [14]. C-CBT or i-CBT can lower barriers to seek help, which is especially important
82 in settings such as those in sub-Saharan Africa with low levels of access to psychologists and therapists
83 [15].

84 C-CBT and i-CBT have evolved into interactive, automated conversational agents (chatbots) that are
85 driven by artificial intelligence, including the recent entrant Vitalk (with versions available for the public
86 and adapted for health professionals). A panel study without control group in Brazil established the
87 effectiveness of the public version, finding a large post-intervention effect size of Cohen's d of -0.81 or
88 greater for anxiety, depression and stress [16]. The version for health workers has been pilot tested in
89 Malawi, a country in which access to mental health therapy is very limited, with only 0.02 psychologists,
90 0.01 psychiatrists and 0.04 occupational therapists per 100,000 population [17]. Only recently has the
91 country successfully trained 3 psychiatrists. The frequent deployment of mental health professionals,
92 especially psychiatric nurses, to other duties such as maternity services exacerbates this situation [18].

To date, there have been few studies, and even fewer RCTs, with the aim of establishing the effectiveness of chatbots. Most of these trials were based on small samples of 70 participants or fewer, had a short duration of 2-4 weeks, suffered from serious biases, and fell short of establishing that chatbots lead to improved mental health outcomes. All trials were conducted in high-income countries, and none in low-resource settings [19]. No RCTs have been conducted to assess the effectiveness of chatbots in improving the mental health status of health workers specifically [15].

To fill these gaps in evidence, we conducted an RCT in Malawi (building on the earlier pilot test), to investigate our working hypothesis that a virtual mental healthcare assistant chatbot, Vitalk, is an acceptable source of psychosocial and mental wellbeing support for health workers to effectively decrease work-related anxiety, depression, burnout, and loneliness (based on standard mental health scales), and to increase resilience and resilience-building behaviors.

Methods

Study design

This study, a 2-arm, parallel RCT, with a pre-treatment assessment, 8-week intervention period, mid-study assessments at 4-5 weeks, and an end-of-study assessment at 9 weeks, was conducted by 2 study teams with distinct roles: a trial management team and a research team. The management team reviewed registration data, removed exclusions, assigned groups and unique trial IDs, and provided communications and technical support to participants. The research team developed the study protocol, recruited participants, held participant workshops at the beginning and end of the study, held focus group discussions (FGDs) during the final workshop, and analyzed the data (de-identified before the analyses).

114 Participants were randomly assigned to either the treatment or control group after registration. Vitalk,
115 the developer of the app, created a web portal for participants to register, give consent, and enter their
116 demographic information. Participants received an email with their unique trial ID and a link to either
117 the treatment app, Vitalk, or to the website (control group). As a single-blinded design, only the
118 research team was blinded to the study arm assignments; participants in both arms were told they were
119 participating in a study on "online self-help for mental wellbeing" but they were aware of the specific
120 intervention they were using and may have found out that others used a different intervention. Neither
121 personally identifiable information nor the participants' facility affiliations were shared outside the trial
122 management team.

123 Study population and recruitment

124 The 8 professional cadres (Table 1) eligible for participation were recruited from all public and private
125 primary, secondary and tertiary care facilities within Blantyre and Lilongwe districts.

126 **Table 1.Type of health workers eligible.**

• Doctors	• Nurses
• Medical Assistants	• Clinical officers
• Laboratory technicians	• Physiotherapy technicians
• Pharmacists	• Physiotherapists

127

128 We recruited participants via posters placed in participating health facilities. All facilities were contacted
129 by phone and email; in larger facilities the research team made short presentations to leadership to seek
130 permission to recruit from their facilities and hang the posters.

Informed consent, screening and eligibility

Interested health workers were invited to a half-day workshop to provide general information and answer questions about the study, and in which they could choose to join the study by signing an informed consent form (kept under lock by the University of Malawi). The research team introduced the trial as a "mobile-phone-based research study of online self-help for mental wellbeing" but provided no details about the 2 study groups.

To be eligible, participants needed to meet the following inclusion criteria:

- Be currently employed as 1 of the 8 types of service providers listed above (a minimum educational qualification of a diploma adhering to the Ministry of Health employment criteria; medical assistants with a certificate and 2 years of college completed qualified)
- Possess some degree of English language proficiency
- Own a smartphone with an Android operating system
- Have no history – past or current – of counseling or therapy for severe mental health disorders
- Not have self-reported suicidal ideation (question 9 of the PHQ-9)
- Score below "very high risk" levels for depression, or below "very high risk" of anxiety combined with "high" or "very high risk" of depression on the initial and subsequent assessments

The management team emailed excluded participants Internet resources for self-help. For those excluded on the basis of suicidal ideation or very high risk of depression and/or anxiety, the team also sent the contacts of 2 local psychologists.

Study procedures and participant retention

The University Research Co. (URC) Institutional Review Board and the University of Malawi Research Ethics Committee (UNIMAREC) approved the study protocol.

In Malawi, Internet access and the use of apps through mobile phones requires expensive data packages. Participants received 3 data packages of 5 gigabytes (GB) valued at $10 each (distributed during the startup workshop, halfway through the study, and at the final workshop). This data allowance covered a minimum of at least 1 month of an hour of daily use of Vitalk or web resources At the startup and final workshops, participants received a transport allowance ($10) and a food allowance ($5), following government guidelines. These allowances were not tied to active participation in the trial to avoid any undue influence, and participants received no additional incentives. As the principal means of participant retention, both study groups received weekly email and WhatsApp messages encouraging them to engage with their app or website resources. During the baseline period, the management team provided the treatment group with detailed download and sign-in instructions for the app via WhatsApp, after noticing low sign-in and receiving messages about sign-in difficulties. The research team was blinded to this process.

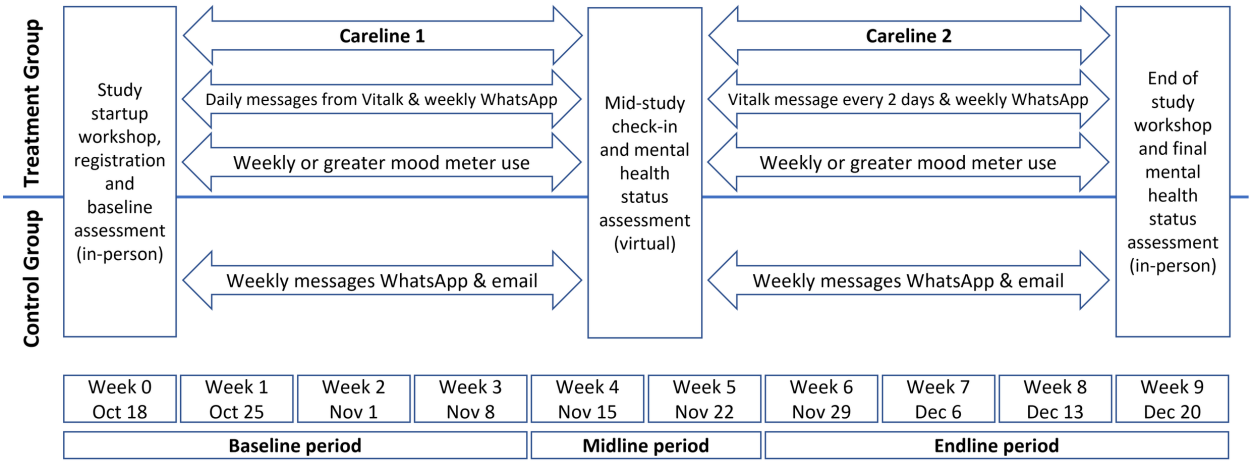
Study timeline

The study spanned 56 days (October 25-December 19, 2021), with startup workshops the week of October 18 and end-of-study workshops the week of December 20 (Fig 1). The research team held several workshops each time to accommodate the large number of participants and their varied work schedules, while complying with physical distancing requirements due to COVID-19. The length of the study enabled treatment group participants to complete exercises for 2 “carelines” (thematic conversations addressing anxiety, depression, stress, relationships, or the COVID-19 pandemic) over a period of a month for each. The baseline, midline, and endline periods take into account that

174 participants completed their first mental health assessments at varying times after registration and

175 repeated assessments with varying frequencies before completing their selected carelines. Several

176 participants completed assessments in a post-endline period.



177 **Fig 1. 2021 Study timeline.**

178 Sample size and randomization

179 The required study population size for paired tests of correlated means was calculated using STATA 17

180 and is based on the following information from published literature:

181 - Pre-post-treatment difference of 1.5 points on standard test scales
182 - Standard deviation of the difference = 6
183 - Power 80%
184 - Significance level 5%

185 This resulted in a minimum of 128 participants per study arm. Given the continuity experience (the

186 proportion of participants completing the pre- and post-treatment assessments for depression, anxiety

187 and stress) of 20-45% reported by Daley et al. [16], we assumed that the dropout rate could be as high

188 as 75% for at least 1 of the standard mental health assessments. This required a sample of 512

189 participants per arm to yield an effective post-intervention sample of 128. Furthermore, if 20% of

190 potential participants did not meet inclusion criteria or dropped out for other reasons, about 640
191 enrollees were needed initially per arm.

192 The management team assigned randomized trial participant numbers (trial ID) using permuted block
193 randomization sequences, generated by the website Randomization.com [20], to maintain a balance
194 across the study arms. Block size varied between 4 and 10 randomly ordered treatment assignments. A
195 random sequence of trial IDs between 1001 and 3000 was generated using the Research Randomizer
196 [21]. Using randomized trial IDs and permuted block randomization ensured that only the management
197 team was able to identify participants' group assignment and that the research team remained blinded
198 throughout trial implementation.

199 Study interventions

200 Treatment group

201 Participants in the treatment group were asked to download and register with the mental health
202 chatbot app, Vitalk, and to use it daily over the 8-week study period. The aim of Vitalk is to improve
203 mental wellbeing by reducing stress, anxiety, depression, and burnout and increasing resilience using a
204 preventative approach to mental health. Vitalk is an automated chatbot, using an avatar named Viki (Fig
205 2) to deliver mental health content in an innovative conversational format.

Fig 2. Viki, Vitalk's virtual mental healthcare assistant.

The conversations are based on insights and strategies taken from Cognitive Behavioral Therapy (CBT) and Positive Psychology [22, 23], which include (but are not limited to): psychoeducation; cognitive restructuring; behavioral activation; gratitude; and practical exercises (breathing, relaxation, and meditation). The goal of the conversations is to help users reflect on experiences and learn techniques to manage stress, mood, and anxiety. A mood tracking tool aids reflection, "emojis" mimic a natural human interaction, and gamification increases user engagement [24]. Vitalk users choose between different thematic conversations ("carelines") and use standardized instruments to assess their mental health and resilience status before and after completing a careline at monthly or more frequent intervals. The app provides users with feedback and advice about their assessment scores. Carelines are continuously refined to ensure the language and emotional expression matches the needs of different users. If user input in the PHQ-9 questionnaire suggests suicidality (e.g., "life isn't worth living") or suicidal ideation, the risk module is triggered and the chatbot delivers information about emergency support, including the contact information of local psychologists.

The version used in Malawi consisted of 5 carelines: anxiety; depression; stress; relationships; and the COVID-19 pandemic. Throughout the 56-day study period, users selected 2 carelines to complete (4 weeks for each). During the first careline period (the first 4 weeks of the study), the user was engaged in

223 daily conversations, each lasting 5-10 min. This changed to an engagement every other day during the
224 second careline period (4 weeks). After the study, users were free to continue using the app and
225 experience the other carelines, receiving a final data bundle at the last workshop to enable them to do
226 so.

227 The chatbot avatar Viki was developed through focus groups and pilot testing in Brazil. A team of clinical
228 psychologists and healthcare professionals from Vitalk wrote the conversations used by Viki and
229 adapted them to the Malawi context. Unlike that of Brazil, the Malawi version did not include natural
230 language understanding (NLU). The chatbot was styled and designed by an information technology,
231 product development and user experience (UX) team at Vitalk in São Paulo. The app is built on ruby and
232 JavaScript as a free-to-use service, hosted within an instant messenger platform, accessible from
233 Android devices.

234 Vitalk does not aim to replace a healthcare professional or to offer treatment. Users are made aware of
235 this limitation in the terms and conditions required for consent; they are advised to seek additional
236 support if they show a high risk of depression, anxiety or burnout in their mental health assessments.

237 Control group (wait-list)

238 Participants in the control group were provided access to a webpage in their Internet browser with links
239 to 4 mental health resources: Mental Health Foundation in the United Kingdom[25]; WHO/Geneva
240 mental well-being resources for the public [26]; WHO/Geneva #HealthyAtHome - Mental health [27];
241 and the Doing What Matters in Times of Stress: An illustrated Guide [28]. These resources had no
242 interactive features (passive) and relied on participants' initiative to access them. Control group
243 participants were considered to be wait-listed since they were given access to the Vitalk app after the
244 completion of the study.

245 # Outcomes

246 ## Primary outcome measures

247 This study used the scores from several standardized mental health and resilience assessments as
248 outcome measures. In addition, we introduced a new measure of resilience-building activities. In the
249 analysis, we used total assessment scores as well as risk groupings to establish clinical relevance
250 following classifications from published literature (based on the sensitivity and specificity of the
251 assessments).

252 **Patient Health Questionnaire (PHQ-9)**

253 The PHQ-9 is a 9-item self-reported scale that evaluates symptoms of depression over the past 2 weeks
254 (e.g., "how often have you been bothered by feeling down, depressed, or hopeless"). Item response
255 options use a Likert scale ranging from 0 (not at all) to 3 (nearly every day), with a maximum score of 27.
256 According to the literature, total scores are divided into 5 risk categories: none (0–4); mild (5–9);
257 moderate (10–14); moderately severe (15–19); and severe (20+) symptoms. The PHQ-9 has been widely
258 used and validated in Malawi [29].

259 **Oldenburg Burnout Inventory (OLBI)**

260 Burnout is linked to relatively high work requirements and limited resource availability for managing
261 them. The discrepancy between resources and challenges creates a significant negative emotional state.
262 The OLBI has 16 items rated 1 to 4, with a maximum score of 64. Eight items describe exhaustion (OLBI-
263 E) and 8 describe disengagement (OLBI-D) [30]. The questionnaire includes both straight and reversely
264 worded items in both dimensions. According to the literature, total scores are divided into 3 risk
265 categories: low (16-35); moderate (36-43); and high (44+). OLBI has been validated and used in low- and
266 middle-income countries (LMICs) [31].

267 **Generalized Anxiety Disorder (GAD-7)**

268 The GAD-7 is a 7-item self-reported scale to assess anxiety symptoms over the past 2 weeks (e.g., "how
269 often have you been bothered by feeling afraid something awful might happen"). Scores range from 0
270 (not at all) to 3 (nearly every day), with a maximum score of 21. The total scores are divided into 4 risk
271 categories: none (0–4); mild (5–9); moderate (10–14); and severe (15+) symptoms. GAD-7 has been
272 utilized effectively in Malawi and other LMICs with comparable demographics [32].

273 **UCLA Short (three-item) Loneliness Scale (UCLA Loneliness)**

274 UCLA Loneliness was added to Vitalk for this study. While there is no agreed upon definition of
275 "loneliness" in research, it is described as an unwelcome, painful and unpleasant feeling, and a fluid
276 experience that can come and go over a short time or persist in the longer term. The UCLA Loneliness
277 scale consists of 3 items rated 1 to 3, with a maximum score of 9. Total scores are divided into 2 risk
278 categories: non-lonely (3-5) and lonely (6+) [33, 34].

279 **5-item Resilience-Building Behavior Scale**

280 A novel addition to this study and Vitalk was an assessment of participants' resilience-building behaviors
281 related to stress management, self-awareness, self-care, purpose, and connection with others (adapted
282 from a toolkit for building health worker resilience [35]). While numerous scales for measuring resilience
283 exist, we did not find any publications on the validity and reliability of instruments measuring resilience-
284 building activities. The questionnaire had 5 items rated 1 to 4, with a maximum score of 20, related to
285 the frequency with which participants performed resilience-building activities over the past 2 weeks.
286 Scores were divided into 3 activity levels: low or none (5-10); moderate (11-15); and high (16-20).

287 **14-item Resilience Scale (RS–14)**

288 Resilience refers to the ability to withstand or adaptively recover from stressors, promoting
289 psychological and physical wellbeing. Resilience is negatively correlated with symptoms of generalized
290 anxiety and post-traumatic stress and positively correlated with gratitude, optimism, and positive affect.
291 The 5 characteristics of resilience are: meaningful and purposeful life; perseverance; equanimity; self-
292 reliance; and existential aloneness [36]. The RS-14 has 14 items rated 1 to 7, with a maximum score of
293 98. Total scores are divided into 6 resilience levels: very low (14–56); low (57–64); on the low end (65–
294 73); moderate (74–81); moderately high (82–90); and high (91+). RS-14 has been validated and used in
295 LMICs [37, 38]

296 While the app registers the actual scores, to simplify communication of risk to users, Vitalk reports 4
297 levels of mental health risks and resilience for its 4 standard assessments (GAD-7, PHQ-9, OLBI, and RS-
298 14): very low; low; high; and very high risk. The Vitalk risk classification differs from published levels as
299 follows: 1 additional level for OLBI; 4 instead of 5 levels for PHQ-9; and 4 instead of 6 levels for RS-14.
300 For the purposes of this trial, we used only published risk classifications based on actual test scores.

301 All 6 assessments together required about 20 minutes to complete. Participants in both arms were
302 allowed to take the assessments repeatedly and saw their results immediately after answering the 6
303 questionnaires. Participants were prompted either by the avatar Viki or What's App messages and
304 emails to take the assessments at specific intervals during the baseline, midline, and endline periods.

305 Secondary measures

306 Mood meter

307 Users of the Vitalk app had the option of assessing their mood using emojis representing 5 mood levels
308 (0-4): very bad; bad; okay; good; and very good. Emojis are successfully used in sentiment analysis and
309 show a stronger emotion compared to words [24]. The timing of mood assessments was matched to

310 mental health assessments prior to data analysis since participants took them independently. The mood
311 meter was part of the intervention to raise self-awareness, and only available to the treatment group.

312 **Participant engagement**

313 For the treatment group, we captured several measures related to participant engagement over the
314 study period: number of interactions with Vitalk; number of days with interactions with the app; number
315 of days between the first and last day of interaction with the app; and total number of hours spent on
316 the app. For the control group, we captured 2 measures related to participant engagement: number of
317 days they accessed web resources and number of web resource click-throughs. All raw engagement
318 measures were converted to z-scores for both groups to account for differences in metrics and
319 distributions in each group. Standardized hours spent on Vitalk, and standardized number of days with
320 web resources accessed, were retained as determinants of mental health and resilience scores.

321 **Smartphone experience**

322 During registration, we measured participants' smartphone experience by asking how often (never, less
323 often, weekly, several times a week, daily) they used the following features (each scored 0 to 4):
324 Internet; WhatsApp; Facebook, Instagram or twitter; chat app; mobile money; games; email; videos; and
325 music. Smartphone experience is a composite of the frequency of use for all 9 features, with a maximum
326 score of 36.

327 **Current or past exposure to mental health counseling**

328 During registration, participants were asked whether they had any current or past exposure to mental
329 health counseling. Possible responses were: currently receiving counseling; received counseling in the
330 last 6 months; received counseling in the last 2 years; received counseling more than 2 years ago; and
331 never received counseling. Participants with any counseling experience were then asked whether

332 counseling was for mild or moderate mental health issues, or for acute or severe mental health issues.
333 Participants who responded the latter were excluded from the study and given the contact for local
334 psychologists.

335 **Current or past chat app and website use for mental health**

336 At registration, participants were asked about their current use of a chat app for mental wellbeing (with
337 the response options: never; sometimes; and often). Past use was assessed through a binary yes or no
338 response. We created a composite "chat app for mental health" variable with values ranging from 0 (no
339 chat app use) to 5 (current often use and past use) by combining responses from both questions.
340 Current and prior use of Internet-based resources about mental wellbeing was assessed in the same
341 manner.

342 **Impact of COVID-19**

343 Four questions assessed how the COVID-19 pandemic impacted study participants. The first 3 questions
344 asked about the impact on workload, workhours, and stress levels over the past 12 months. Possible
345 response options for each question (scored from 0 to 4) were: greatly decreased; somewhat decreased;
346 stayed the same; somewhat increased; and greatly increased. The sum of the responses to these
347 questions created a composite COVID-19 effect variable. The fourth question inquired whether COVID-
348 19 prevented participants from going into work over the past 12 months. Response options (scored from
349 0 to 3) were: no disruptions at all; several days over the entire year; several days every month; several
350 days every week; and I did not work at all (coded as missing).

351 **Anonymous online participant experience questionnaire**

352 During the final workshops, participants from both arms had the opportunity to complete an
353 anonymous online questionnaire about their experience using either the Vitalk app or the web

354 resources. Questions focused on the frequency of use, reasons for infrequent use, relevance and
355 usefulness, ease of use, and specific difficulties encountered. Some key findings related to study
356 outcomes are highlighted in this paper; a separate comprehensive analysis will be published together
357 with the findings from focus group discussions.

358 Data analysis

359 Participant characteristics are presented as proportions or means (SD) for the treatment and control
360 groups.

361 Mixed linear model. We used multilevel mixed-effects generalized linear models (STATA *meglm*
362 command) to analyze primary participant-reported outcome measures for all 6 mental health and
363 resilience assessments over the baseline and endline periods. A full factorial model of study group and
364 assessment period was fitted to represent fixed effects. Participants' trial ID specified the random effect
365 in this model. The parameter estimate of interest was the coefficient of the interaction between study
366 group and assessment period, which is equivalent to a difference-in-differences (DiD) estimator. DiD
367 estimators were robust, because only valid assessments for the baseline and endline periods were
368 included in the analysis, avoiding the need for inter- or extrapolating missing values. Because of
369 randomization, covariates were not included in the mixed models. Data from the midline and any
370 assessments completed after the 8-week study period (post-endline) were not included in the analysis
371 because of a much lower number of observations in these compared to the other periods.

372 Effect size. To compare the different outcome measures, Cohen's d was calculated as the mean
373 difference between the 2 groups divided by their common SD and interpreted as small (d=0.2), medium
374 (d=0.5), and large (d=0.8).

375 Reliable change index. To assess whether the change in outcomes (the total score for each assessment,
376 between baseline and endline) is statistically reliable, and not due to measurement variability inherent
377 to the assessments alone, we calculated a reliable change index (RCI). The RCI is the standard error of
378 the difference between the 2 measurements, SE_{diff}, in equation (1). SD_{bas} is the standard deviation of the
379 baseline observations, and r is Cronbach's alpha representing the reliability of the measurement
380 instrument. Each assessment instrument has its own RCI.

381 $$RCI = 1.96 \times SE_{diff} = 1.96 \times SD_{bas}\sqrt{2}\sqrt{1-r} \quad (1)$$

382 Change in risk levels. We used total scores from the 4 mental health and 2 resilience assessments for the
383 mixed linear model, effect size, and RCI estimation. To determine whether the changes between the
384 pre- and post-intervention periods matter from a clinical perspective, the total assessment scores – a
385 continuous variable – were converted to mental health risk and resilience levels –an ordinal variable –
386 using the published classifications described above for each type of assessment (e.g., depression scores
387 from the PHQ-9 are divided into 5 levels ranging from none to severe). A change in risk levels was
388 calculated as either a drop or increase of 1 or more levels for each mental health and resilience metric.
389 For participants registering a change in risk level, the proportion with reliable change was calculated.
390 Reliable change was established if the difference between baseline and endline scores were equal or
391 greater than the RCI.

392 Determinants of mental health and resilience. All mental health and resilience outcomes were regressed
393 against all participant characteristics using ordinary least squares models, controlling for study group
394 and assessment period. All observations with valid baseline and endline assessments for the control and
395 treatment groups were included in the analysis. A separate analysis was done for the treatment group
396 to assess which covariates unique to this group (mood meter scores and carelines) affected the
397 outcomes.

398 $P \leq 0.10$ was considered to be indicative of statistical significance. Equally important as the p-value was

399 the consistency of patterns across all mental health and resilience assessments and their agreement

400 with behavioral theory. All data were analyzed with STATA 17.0 (StataCorp LLC, College Station, TX).

401

Results

The Consolidated Standards of Reporting Trials (CONSORT) diagram in Fig 3 shows that of 1,584 participants enrolled in the study, 836 completed one or more mental health and resilience assessments (481 in the control and 355 in the treatment group). Based on participant feedback, the treatment group saw a much greater attrition due to technical difficulties impacting the use of the Vitalk app. The number of participants completing assessments varied substantially by study period, as seen in Table 2.

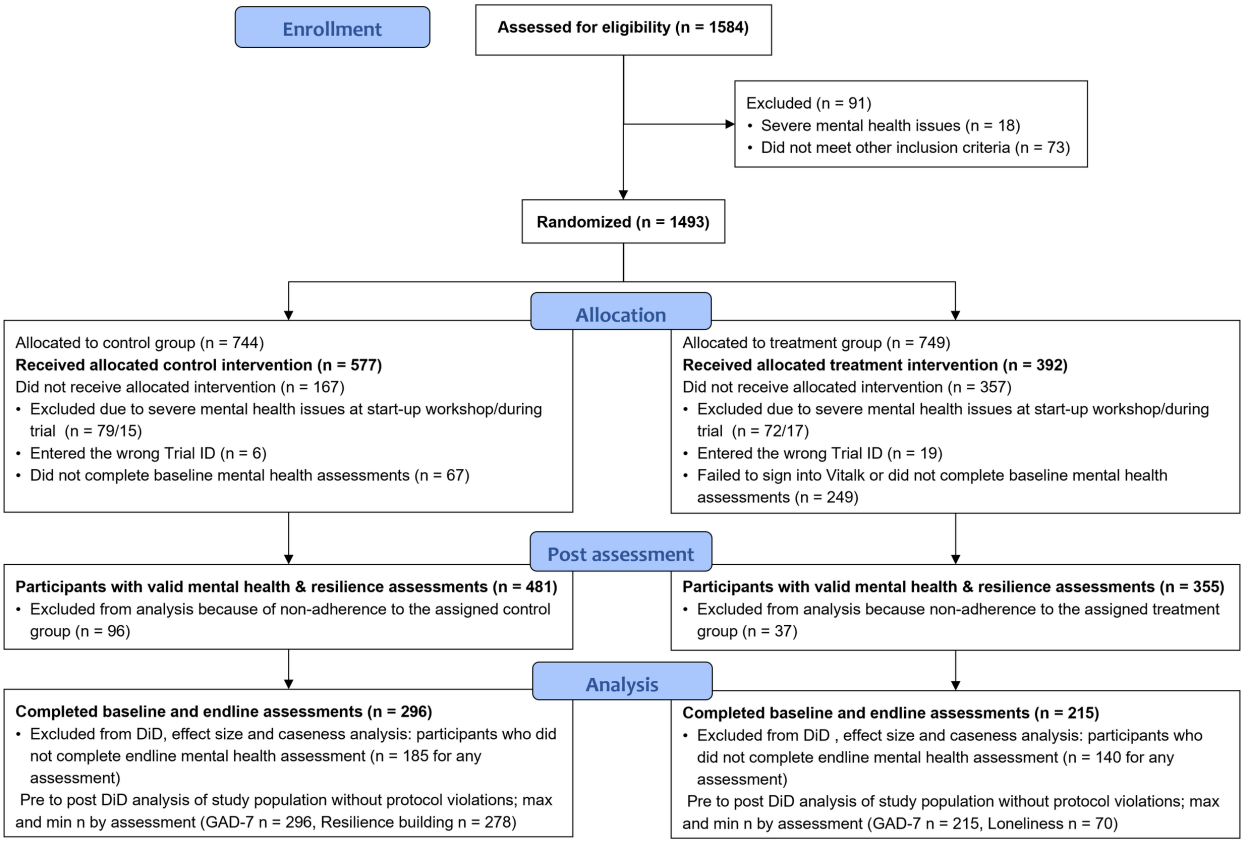


Fig 3. Study participant flow.

410 **Table 2. Number of participants completing any mental health and resilience assessment by study**
411 **period.**

Study period	Control Group	Treatment Group	Total
Baseline	469	318	787
Midline	77	181	258
Endline	304	260	564
Post-endline	13	86	99

412 Numbers include singular assessments that were completed in one study period only, which
413 were excluded from analysis.

414

415 For the DiD, effect size, and risk level analyses, Table 3 shows that the sample size differed by type of
416 assessment. Overall, 511 participants (296 in the control and 215 in the treatment group) had no
417 protocol violations and completed baseline and endline assessments for GAD-7. Only participant who
418 completed base- and midline assessments were included in the calculations of DiD estimators using
419 mixed-effects linear modeling, effect size estimation, and change in risk levels. Samples in the treatment
420 group for the loneliness and resilience-building activities assessments were markedly smaller than in the
421 control group. These 2 assessments were the last in the series of the 6 assessments suggesting that a
422 majority of Vitalk users stopped completing assessments after RS-14. The control group did not
423 experience a similar drop, suggesting that technical issues may have prevented app users from
424 completing all assessments.

425 **Table 3. Effective sample size by type of mental health and resilience assessment with completed**
426 **baseline and endline.**

Assessment	Control Group n	Treatment Group n	Total n
Anxiety (GAD-7)	296	215	511
Depression (PHQ-9)	286	201	487
Burnout (OLBI)	291	211	502
Resilience (RS-14)	280	183	463
Loneliness (UCLA)	280	70	350
Resilience building activities	278	85	363

427 Assessments shown in the order completed by participants.

428 These are the sample sizes used in DiD, effect size, and change in risk level analyses.

429 ## Participant characteristics

430 The distribution of characteristics for the 836 participants with valid mental health and resilience
431 assessments was balanced between the control and treatment groups as shown in Fig 4. The difference
432 between groups varied by 1 to 6 percentage points. Overall, participants were young (under 30 years)
433 and female. Most were professional nurses working in public secondary or tertiary healthcare facilities
434 in urban areas. Maternity care was the most common type of care provided. Fewer than 30% had ever
435 used a chat app for mental health and more than 40% had ever accessed websites with mental health
436 content. One in 6 participants had a history of mental health counseling. The level of experience in
437 performing 9 different tasks with their smartphones was high (a mean of 29 out of a maximum score of
438 36) and nearly the same in both study groups (Fig 5), corresponding to several times per week to daily
439 use. Both groups reported the same effect of the COVID-19 pandemic. More than 1 in 3 participants
440 mentioned that the pandemic had disrupted their work. The average effect of COVID-19 on participants'
441 workload, workhours, and stress levels was 8 out of a maximum score of 12 (Fig 5), corresponding to
442 somewhat increased workload and stress levels.

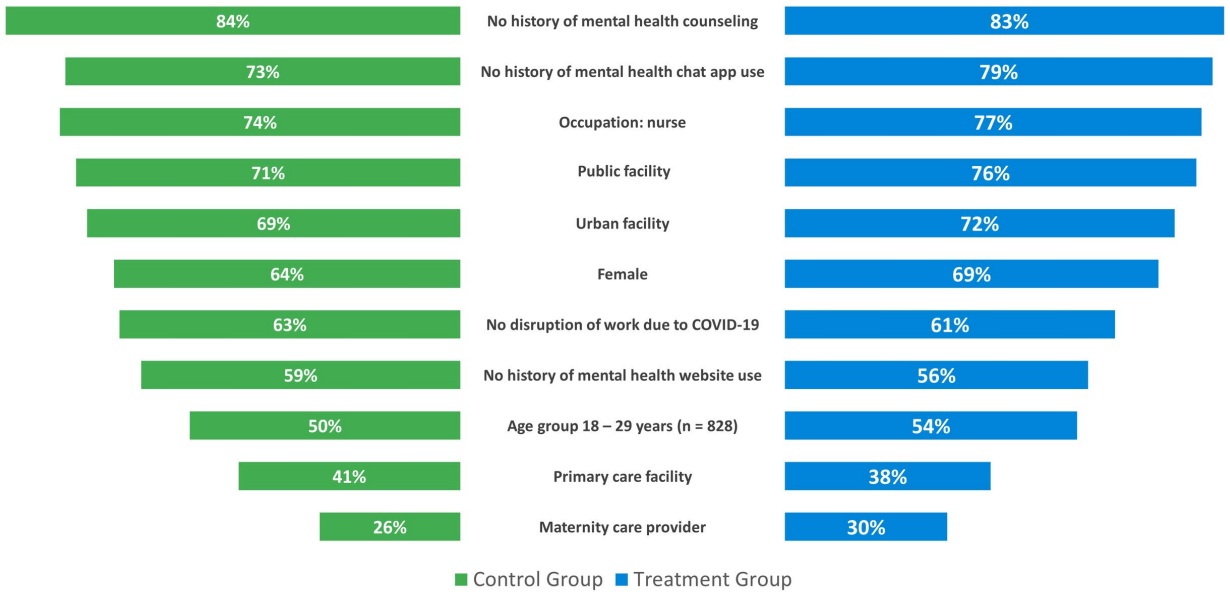


Fig 4. Study participant characteristics. Only the most frequent categories are shown (n = 836 [481 control group, 355 treatment group] except Age)

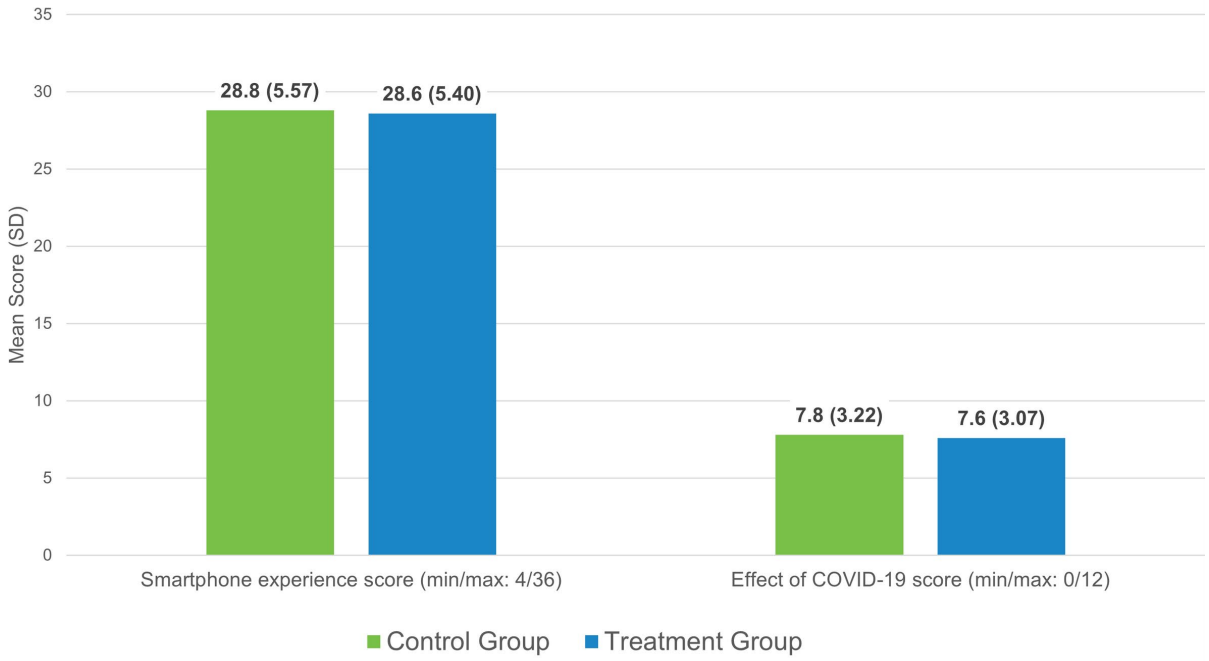


Fig 5. Mean smartphone experience and COVID-19 impact scores (n = 836).

Participant adherence to treatment and control interventions

Treatment group participants had an average of 2,200 (SD 6,294, median 920, range 84-58,021) interactions with the Vitalk app over the 56-day study period, with a distribution skewed towards the lower end. The average duration of app use was 24 days (SD 14, median 23, range 1-67), and users spent an average of 10.4 hours (SD 7.7, median 9.2, range 0.4-51.7) on the app. Treatment group participants expressed their mood an average of 13 times (SD 15, median 8, range 0-81) over the study period. The mean mood score was 2.9 (SD 1.0, median 3.0, range 0-4).

Control group participants accessed the website resources an average of 1.9 days (SD 2.0, median 1, range 1-23) over the study period, with a distribution heavily skewed towards 1.0 days. The only other measure of adherence available in this group was an average number of 1.2 click-throughs (SD 0.7, median 1, range 1-8) of links to mental health resources on the website. We did not measure time spent on resource use and mood status for the control group.

Summary of outcomes at baseline and endline

Table 4 shows summary statistics for the 4 mental health and 2 resilience assessments at baseline and endline. Mean anxiety, depression, and loneliness scores from the GAD-7 (baseline mean 4.5 [SD 3.8], endline mean 2.9 [SD 3.5]), PHQ-9 (baseline mean 4.0 [SD 4.1], endline mean 2.9 [SD 3.8]) and UCLA loneliness (baseline mean 5.3 [SD 1.5], endline mean 4.7 [SD 1.6]) questionnaires, respectively, fell into the lowest risk level. Burnout scores were at moderate risk, with a mean of 38.4 (SD 6.0) and 36.6 (SD 6.7) at baseline and endline, respectively. Resilience was moderate to moderately high, with a mean of 78.6 (SD 10.4) at baseline and a mean of 81.5 (SD 11.4) at endline, and resilience-building activities were moderate, with a mean of 13.1 (SD 3.0) and a mean of 14.2 (SD 3.3) at endline. Both control and treatment groups saw an improvement in mental health and resilience scores between the 2 assessment periods. All summary statistics are based on participants with valid baseline and endline

469 data using all assessments completed during each assessment period. The average number of

470 assessments taken each period varied between 1 and 5, depending on the type of assessment and was

471 generally higher at endline.

472

473 **Table 4. Mental health and resilience scores from completed assessments by study period and group.**

Questionnaire	Control Group		Treatment Group		Total	
	Baseline	Endline	Baseline	Endline	Baseline	Endline
PHQ-9						
Frequency	510	634	429	496	939	1130
Mean score ***	3.67	2.99	4.29	2.72	3.95	2.87
Standard deviation	4.01	4.04	4.26	3.52	4.13	3.82
Median score	3	1	3	1.5	3	1
Avg. no. of assessments	2.9	3.3	3.2	4.6	3.0	3.9
OLBI						
Frequency	471	583	379	441	850	1024
Mean score ***	38.00	36.55	38.84	36.63	38.37	36.59
Standard deviation	6.18	6.83	5.74	6.41	6.00	6.65
Median score	39	37	39	37	39	37
Avg. no. of assessments	2.6	3.0	2.8	4.3	2.7	3.6
GAD-7						
Frequency	545	669	468	550	1013	1219
Mean score ***	4.24	2.89	4.73	2.93	4.46	2.91
Standard deviation	3.74	3.63	3.86	3.31	3.80	3.49
Median score	4	1	4	2	4	2
Avg. no. of assessments	3.1	3.6	3.3	4.9	3.2	4.2
UCLA Loneliness						
Frequency	433	531	84	106	517	637
Mean score ***	5.18	4.69	5.70	4.92	5.26	4.73
Standard deviation	1.52	1.59	1.45	1.74	1.52	1.61
Median score	5	4	6	5	6	4
Avg. no. of assessments	2.4	2.7	1.3	2.0	2.2	2.6
Resilience Building						
Frequency	422	521	128	117	550	638
Mean score ***	13.20	14.01	12.80	14.84	13.11	14.16
Standard deviation	3.11	3.40	2.62	2.65	3.00	3.29
Median score	13	14	13	15	13	14
Avg. no. of assessments	2.3	2.7	1.8	1.6	2.2	2.5
RS-14						
Frequency	438	540	327	394	765	934
Mean score ***	79.23	81.05	77.69	82.19	78.57	81.53
Standard deviation	10.02	11.36	10.81	10.57	10.39	11.04
Median score	80	83	79	84	80	83
Avg. no. of assessments	2.4	2.8	2.6	4.3	2.5	3.4

474 Mean score ***: A blue frame indicates statistically significant differences between mean scores
475 at baseline and endline based on within group t-tests. P-values are < 0.01 for all assessments in
476 the control and treatment groups.

477 Mean score ***: Shaded cells indicate statistically significant differences between mean scores
478 comparing study periods for each study group based on between group t-tests. P-values are <
479 0.05 for differences at baseline for all assessments and not significant at endline except for
480 Resilience Building where baseline scores did not differ significantly between groups but endline
481 results did.

482 Using published reference levels and the average score from each assessment for each study period,
483 participants were classified according to “moderate to high” mental health risk and “very low to low”
484 resilience levels. As seen in Table 5, fewer participants in the control and treatment groups were
485 classified as “moderate to high” risk for mental wellbeing and “very low to low” resilience levels at
486 endline than at baseline. However, deceases in depression, burnout, and anxiety, and increases in
487 resilience- building activities and resilience were greater in the treatment than control group, consistent
488 with the changes in mean assessment scores shown in Table 4. Loneliness did not follow this pattern,
489 with a larger percentage of participants improving in the control than in the treatment group.

490 **Table 5. Study participants with moderate to high risk levels for mental health or low levels of resilience.**

	Control Group						Treatment Group						Sample at base- and endline		
	Baseline		Endline		Total		Baseline		Endline		Total		Control Group N	Treatment Group N	Total N
Assessment and risk level	n	%	n	%	n	%	n	%	n	%	n	%			
Depression: moderate, moderately severe, severe	32	11%	25	9%	57	10%	30	14%	18	9%	48	11%	291	211	502
Burnout: moderate or high	211	74%	187	65%	398	70%	156	78%	121	60%	277	69%	286	201	487
Anxiety: moderate or severe	37	13%	24	8%	61	10%	29	13%	13	6%	42	10%	296	215	511
Loneliness: lonely	148	53%	105	38%	253	45%	44	63%	36	51%	80	57%	280	70	350
Resilience building: low or no activity	49	18%	44	16%	93	17%	14	16%	2	2%	16	9%	278	85	363
Resilience: very low, low, or on the low end	71	25%	54	19%	125	22%	56	31%	38	21%	94	26%	280	183	463

491 Risk levels are based on the average assessment score per participant for each study period.

492

Difference-in-differences (DiD) estimators

The DiD estimates shown in Fig 6 suggested a significant positive effect of Vitalk in reducing anxiety and depression, and in building resilience and increasing resilience-building activities. The coefficients for the interaction terms between study group and study period showed that Vitalk was more effective than the passive control intervention in reducing depression (DiD estimator -0.68 [95% CI -1.15 to -0.21]), anxiety (DiD estimator -0.44 [95% CI -0.88 to 0.01]), and burnout (DiD estimator -0.58 [95% CI -1.32 to 0.15]). Burnout was just short of statistical significance at the 90% level. Changes in resilience were positive and significantly greater in the treatment group (DiD estimator 1.47 [95% CI 0.05 to 2.88]). Similarly, this group showed significantly greater resilience-building activities (DiD estimator 1.22 [95% CI 0.56 to 1.87]). There seemed to exist no treatment effect on loneliness, which was omitted from the estimation of the RCI and risk level changes.

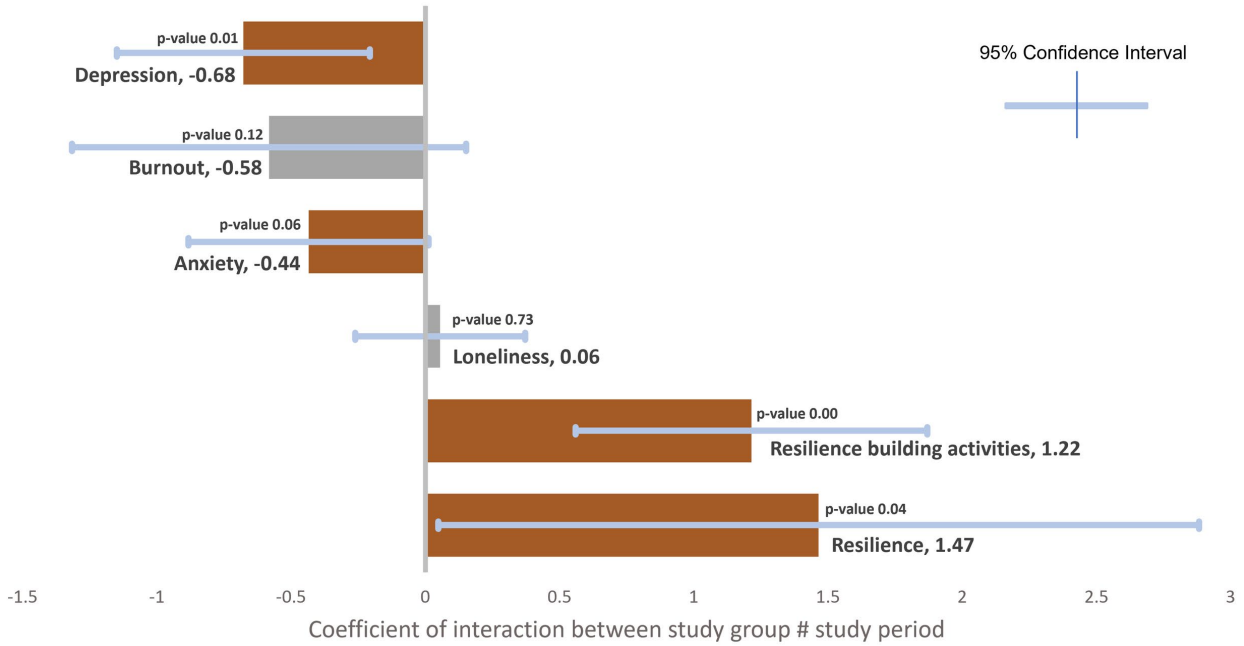


Fig 6. Difference-in-differences estimators with 95% confidence intervals from mixed effects linear models.

506 ## Effect size

507 Although the DiD estimates indicated that Vitalk use positively impacts mental health and resilience
508 outcomes, they did not inform about the size of the effect since each outcome assessment uses a
509 different scale. The effect size for each outcome was measured using standardized mean differences
510 (SMD, Cohen's d). We also calculated Hedges' g – another SMD measure – to account for unequal
511 sample sizes but do not show them separately since the values for 2 decimals were identical to Cohen's
512 d. Fig 7 shows the results comparing the effect size for the control and treatment groups. Participants in
513 the treatment group experienced a consistently larger effect for the mental health and resilience
514 assessments (a medium effect), with the control group trending towards a small effect. Vitalk use had
515 the largest impact on resilience-building activities with a large effect size (more than 3 times the effect
516 in the control group). Effect size is calculated from statistical inference and therefore has confidence
517 intervals (CI) shown in Fig 7. None of the 95% CI contain zero, indicating that both study groups
518 experienced a positive effect – a decrease in depression, burnout, anxiety, and loneliness, and an
519 increase in resilience-building activities and resilience. Depression and resilience-building activities
520 showed the largest difference in effect size between the control and treatment groups, with no
521 overlapping 95% CI.

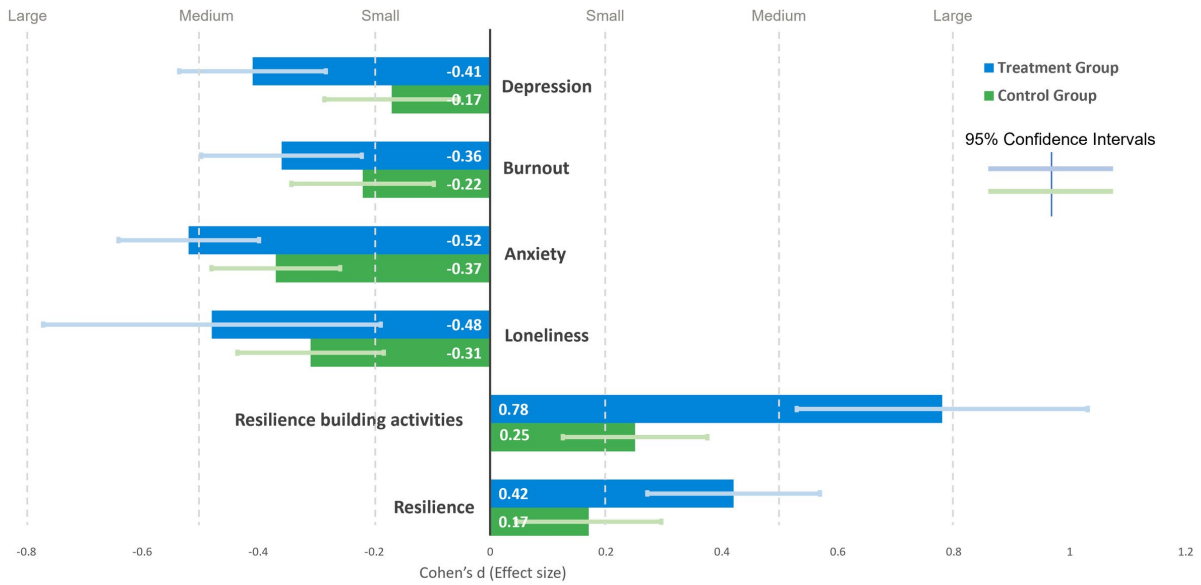


Fig 7. Effect size post-intervention by assessment type and study group.

Reliable change index (RCI)

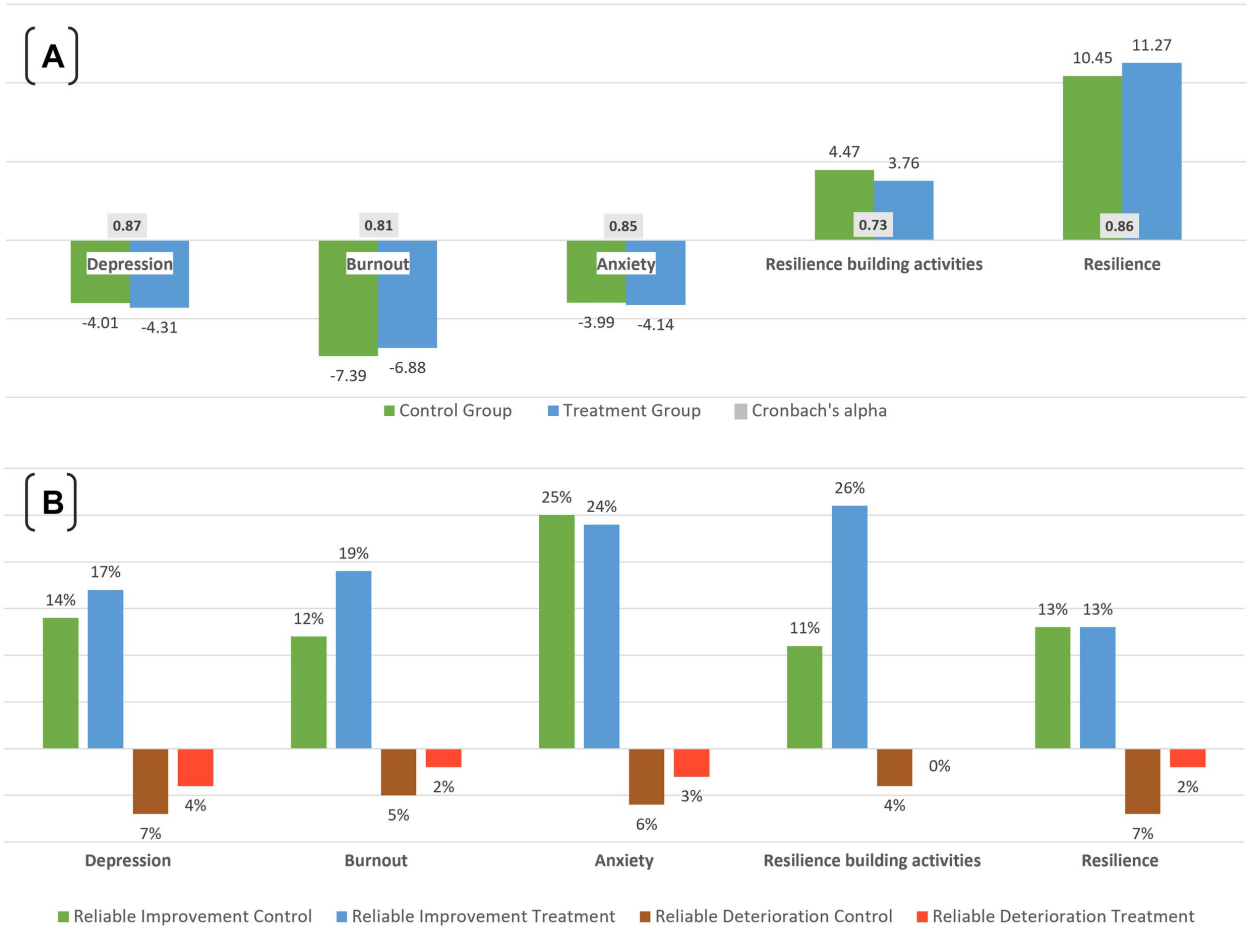
A comparison of the mean and median scores for mental health and resilience outcomes pre- and post-intervention in Table 4 and Table 5 suggests that participants saw improved mental wellbeing and resilience outcomes. However, this is the net result of a larger number of participants improving in outcomes than participants deteriorating in outcomes. To assess whether this change may have arisen by chance alone, we calculated the RCI and the proportion of participants with a change equal or greater than the RCI. The results are shown in Fig 8A and Fig 8B. The RCI for each outcome was similar in the control and treatment groups. Reliable improvement was greater in the treatment than in the control group for depression and burnout, and was about the same for anxiety. The proportion that saw a reliable increase in resilience-building activities was more than twice in the treatment than that of the control group. Reliable change in resilience was the same in both groups. While overall, fewer participants experienced a deterioration in mental health and resilience, reliable change was consistently greater in the control than in the treatment group. Fig 8A also shows Cronbach's alpha as a reliability measure for each assessment, which was used in the calculation of the RCI. Cronbach's alpha

537 is very good (greater and 0.80) for 4 assessment instruments and acceptable for resilience-building

538 activities.

539 **Fig 8. Reliable change in mental wellbeing and resilience outcomes.** [A] Reliable change index (RCI). [B]

540 Reliable improvements or deterioration in outcomes between baseline and endline.



541 ## Mental health and resilience risk level change

542 While any improvement in mean scores for mental health and resilience assessments between baseline

543 and endline is a positive outcome, a difference in means, although significant, did not impart whether

544 these were of clinical relevance. To assess whether a change in scores equated to a change in risk levels,

545 we categorized each participant's scores into risk levels using published cutoff points. We then

546 estimated the proportion of participants changing risk categories by either moving to a lower or higher

547 level (Fig 9). The improvement in risk levels was consistently greater in the treatment than in the control
548 group. The effect on anxiety was greater than on depression or burnout. The depression scale showed
549 the least amount of change, with about two-thirds of participants staying at the same risk level pre- and
550 post-intervention. The proportion of participants increasing resilience-building activities by at least 1
551 level was almost twice as high in the treatment group compared to the control group. The control group
552 consistently saw more cases with a deteriorated outcome than the treatment group. Not all the changes
553 of mental health risk or resilience levels were statistically significant. Therefore, Fig 9 also shows the
554 proportion of all those changing risk levels that were reliable based on the RCI. For burnout, the
555 proportion of reliable change was considerably higher in the treatment group (58%) than in the control
556 group (36%). For anxiety, 71% in the control group saw reliable change compared to 61% in the
557 treatment group. Participants in the control group whose mental wellbeing and resilience outcomes
558 deteriorated between study periods experienced a consistently greater reliable change than those in the
559 treatment group. The proportion of participants with reliable change in Fig 9 is equivalent to the
560 percentage of participants with reliable improvement or deterioration in Fig 8B.

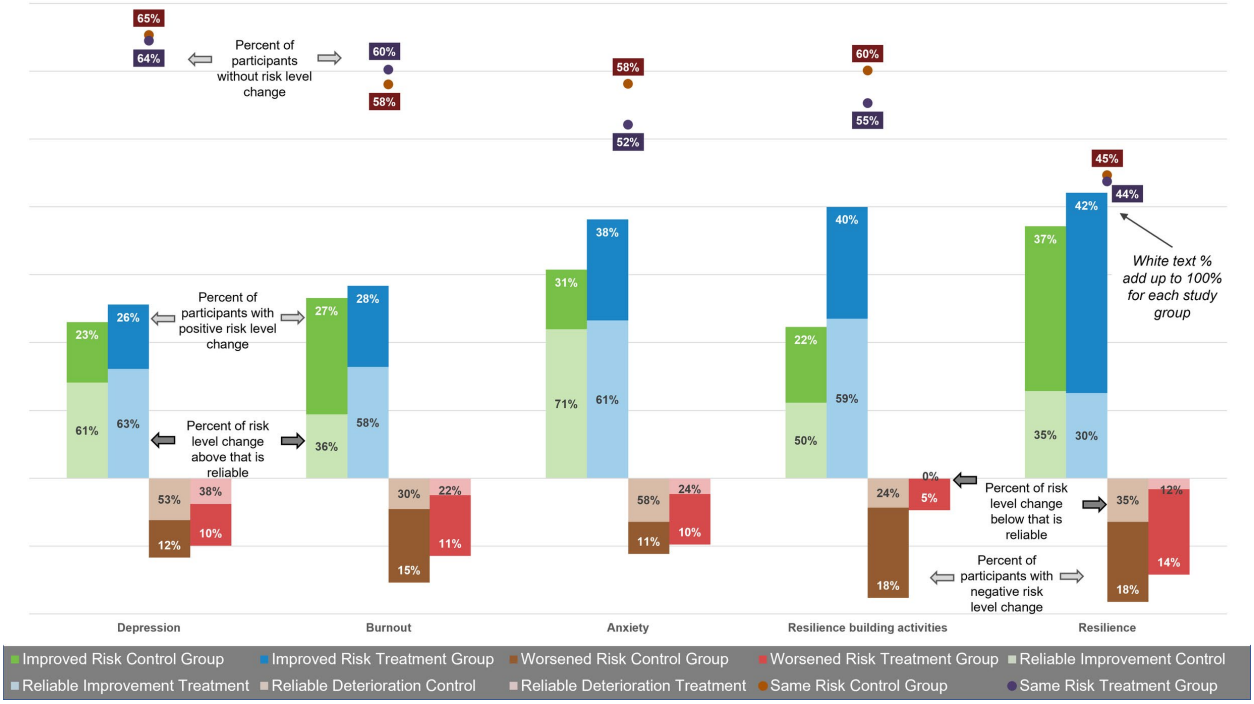


Fig 9. Percent of participants with improved, worsened or same mental wellbeing and resilience risk levels. White color text percent: percent without risk level change or with one or more risk levels difference between baseline and endline. Black color text percent: percent of risk level change that is reliable based on the RCI.

Determinants of mental health and resilience

Ordinary least squares models regressed participant characteristics on mental health and resilience outcomes using data from all participants with valid mental health and resilience assessments, controlling for study group and study period. Table 6 shows that several factors have a significant effect on mental health outcomes. Anxiety increased by 0.02 points (95% CI 0.00 to 0.05) with each year of age. Likewise, depression increased by 0.03 points (95% CI 0.01 to 0.06) and loneliness by 0.02 points (95% CI 0.00 to 0.04). Burnout appeared to decrease with age by -0.04 points (95% CI -0.09 to -0.01). Resilience-building activities increased with age by 0.04 points (95% CI 0.01 to 0.08). Participant gender was negatively associated with anxiety and burnout, with women scoring higher than men by 0.43 points (95% CI 0.07 to 0.79) and 0.90 points (95% CI 0.19 to 1.61), respectively. In addition, women

575 showed less resilience-building activities of -0.40 points (95% CI -0.85 to 0.04) than men. While this was
576 consistent with women having -0.56-point lower resilience levels, this was not statistically significant.
577 Age and gender were also closely correlated with smartphone use experience. With each year increment
578 of age, smartphone experience decreased by -0.17 points [95% CI -0.17 to -0.15]; and females saw -0.99
579 point [95% CI -1.15 to -0.82] lower smartphone experience than men.

580 **Table 6. Determinants of mental health and resilience outcomes for the control and treatment groups combined.**

	Dependent variables (assessment scores)											
	Depression (PHQ-9)		Burnout (OLBI)		Anxiety (GAD-7)		UCLA LONELINESS		Resilience building activities		Resilience (RS-14)	
Independent variables	**Coefficient**	**(SE)**	**Coefficient**	**(SE)**	**Coefficient**	**(SE)**	**Coefficient**	**(SE)**	**Coefficient**	**(SE)**	**Coefficient**	**(SE)**
z-score of time spent @	-0.43***	(0.10)	-0.44**	(0.18)	-0.51***	(0.09)	-0.22***	(0.06)	0.09	(0.12)	-0.04	(0.33)
Cadre: clinical officer †	0.42	(0.36)	1.30**	(0.63)	0.44	(0.32)	-0.17	(0.20)	-0.40	(0.40)	-2.54**	(1.12)
Cadre: doctor †	-0.14	(1.01)	0.65	(1.74)	-0.42	(0.95)	1.37**	(0.56)	-2.74**	(1.17)	-0.94	(3.06)
Cadre: medical assistant †	-1.34***	(0.45)	1.62**	(0.81)	-1.08***	(0.41)	-0.87***	(0.25)	0.57	(0.52)	-3.73***	(1.44)
Cadre: pharmacist †	0.87	(0.89)	2.45	(1.57)	1.14	(0.81)	0.43	(0.43)	0.78	(0.93)	-7.43***	(2.76)
Cadre: technician †	0.82	(0.85)	-1.15	(1.51)	0.93	(0.74)	0.36	(0.55)	-0.16	(1.12)	-1.77	(2.69)
Age (years)	0.03***	(0.01)	-0.04*	(0.02)	0.02**	(0.01)	0.02**	(0.01)	0.04**	(0.02)	-0.02	(0.04)
Sex ‡	0.17	(0.20)	0.90**	(0.36)	0.43**	(0.18)	0.11	(0.11)	-0.40*	(0.23)	-0.56	(0.64)
Care level: secondary $	0.72**	(0.33)	1.46**	(0.59)	0.71**	(0.30)	-0.06	(0.18)	0.37	(0.35)	-0.30	(1.05)
Care level: tertiary $	1.03***	(0.29)	2.01***	(0.52)	0.99***	(0.26)	0.05	(0.17)	0.47	(0.34)	-1.21	(0.93)
Facility type: private nfp ‼	-0.62*	(0.33)	-0.10	(0.59)	-0.60**	(0.30)	0.25	(0.17)	-0.24	(0.34)	-2.08**	(1.04)
Facility type: private fp ‼	0.78*	(0.44)	2.06***	(0.78)	0.48	(0.39)	0.96***	(0.25)	-0.63	(0.49)	-2.48*	(1.37)
Location: peri-urban ∓	0.50	(0.34)	-0.18	(0.61)	0.48	(0.30)	0.01	(0.21)	0.04	(0.41)	-0.68	(1.08)
Location: rural ∓	0.52*	(0.27)	-0.20	(0.49)	0.30	(0.25)	-0.08	(0.15)	0.04	(0.31)	1.50*	(0.88)
Service: counseling §	-1.61***	(0.53)	1.15	(0.96)	-1.01**	(0.48)	0.17	(0.33)	-1.18*	(0.64)	0.33	(1.73)
Service: inpatient care §	-1.20***	(0.41)	-0.04	(0.74)	-1.04***	(0.37)	-0.47**	(0.23)	0.24	(0.47)	-2.44*	(1.32)
Service: intensive care §	-1.40***	(0.45)	0.67	(0.81)	-0.88**	(0.40)	-0.05	(0.26)	-0.21	(0.52)	-0.91	(1.45)
Service: maternity care §	-0.93**	(0.37)	1.46**	(0.66)	-0.56*	(0.33)	-0.42**	(0.20)	0.10	(0.42)	-1.71	(1.18)
Service: support & diagnostics §	-1.69**	(0.84)	-0.73	(1.52)	-1.36*	(0.74)	-0.89*	(0.53)	-0.29	(1.10)	3.69	(2.72)
Service: specialty care §	-2.05***	(0.45)	0.74	(0.81)	-1.59***	(0.41)	-0.76***	(0.25)	-0.53	(0.52)	0.24	(1.43)
Service: surgery care §	-1.55***	(0.48)	0.87	(0.87)	-1.49***	(0.44)	-1.00***	(0.27)	-0.90*	(0.54)	-0.99	(1.55)
Smartphone experience score	0.06***	(0.02)	0.05*	(0.03)	0.08***	(0.01)	0.03***	(0.01)	0.04**	(0.02)	-0.09	(0.05)
History of MH counseling	0.63***	(0.11)	0.35*	(0.20)	0.36***	(0.10)	-0.03	(0.07)	-0.07	(0.16)	0.27	(0.37)
COVID-19 impact score	0.02	(0.03)	0.20***	(0.06)	0.06**	(0.03)	0.01	(0.02)	-0.13***	(0.04)	0.21**	(0.10)
COVID-19 ability to work impact	0.62***	(0.10)	0.95***	(0.17)	0.39***	(0.09)	0.17***	(0.06)	-0.24**	(0.12)	-0.11	(0.31)
History of chat app use for MH	-0.02	(0.08)	-0.24*	(0.14)	-0.07	(0.07)	-0.06	(0.05)	0.28***	(0.09)	0.19	(0.26)

	Dependent variables (assessment scores)											
	Depression (PHQ-9)		Burnout (OLBI)		Anxiety (GAD-7)		UCLA LONELINESS		Resilience building activities		Resilience (RS-14)	
Independent variables	**Coefficient (SE)**		**Coefficient (SE)**		**Coefficient (SE)**		**Coefficient (SE)**		**Coefficient (SE)**		**Coefficient (SE)**	
History of website use for MH	0.20***	(0.08)	-0.07	(0.14)	0.17**	(0.07)	0.05	(0.04)	0.07	(0.09)	0.30	(0.25)
Treatment group ¶	0.42**	(0.18)	0.30	(0.32)	0.47***	(0.16)	0.52***	(0.13)	0.09	(0.24)	-0.45	(0.58)
Study period mid, end, post ≡	-0.60***	(0.09)	-0.95***	(0.15)	-0.80***	(0.08)	-0.27***	(0.05)	0.59***	(0.10)	1.63***	(0.27)

581 P-values: * <= 0.10, ** <= 0.05, *** <= 0.01

582 z-score of time spent @: represents standardized hours spent on Vitalk for the treatment group, and standardized number of days with
583 web resources accessed for the control group

584 Reference categories: † Cadre: nurse, ‡ Male, $ Care level: primary, ‼ Facility type: public, ∓ Location: urban, § Service: outpatient care,
585 ¶ Control group, ≡ Baseline

586 Abbreviations: nfp = not-for-profit, fp = for profit, MH = mental health, mid = Midline, end = Endline, post = Post-endline period

587 Secondary care providers had worse outcomes (higher scores) for anxiety, depression, and burnout
588 compared to primary care providers, ranging from 0.71 points (95% CI 0.12 to 1.30) for anxiety to 1.46
589 points (95% CI 0.30 to 2.62) for burnout. Tertiary care providers scored 2.01 points (95% CI 1.00 to 3.03)
590 higher for burnout than primary care providers. Outpatient care saw consistently higher anxiety,
591 depression, and loneliness than any of the other types of care, with surgery (depression -1.54 points,
592 95% CI -2.50 to -0.60) and specialty care (depression -2.06 points, 95% CI -2.93 to -1.18) having the
593 largest difference. Maternity care was the only type associated with a significantly 1.46-point (95% CI
594 0.16 to 2.76) higher burnout than outpatient care. Medical assistants scored significantly lower for
595 anxiety, depression, and loneliness compared to nurses, but they had 1.62-point (95% CI 0.03 to 3.20)
596 higher burnout, and -3.73-point (95% CI -6.56 to -0.90) lower resilience than the latter.

597 Anxiety, depression, burnout, and loneliness outcomes all decreased with the z-scores of time spent
598 with the treatment or control interventions, ranging from -0.22 points (95% CI -0.33 to -0.10) for
599 loneliness, to -0.51 points (95% CI -0.68 to -0.34) for anxiety. Conversely, greater use of smartphone
600 features was associated with significantly higher levels of anxiety, depression, burnout, and loneliness,
601 ranging from 0.03 points (95% CI 0.01 to 0.06) for loneliness to 0.08 points (95% CI 0.05 to 0.11) for
602 anxiety. Participants with greater smartphone experience had higher resilience-building activities by
603 0.043 points (95% CI 0.001 to 0.084). While smartphone use showed a -0.09-point lower level of
604 resilience, this was not statistically significant (p-value of 0.104).

605 Participants with a history of mental health counseling had higher levels of anxiety (0.36 points, 95% CI
606 0.16 to 0.56), depression (0.63 points, 95% CI 0.41 to 0.85), and burnout (0.35 points, 95% CI -0.05 to
607 0.76). The COVID-19 pandemic and its impact on work-related stress, workload, and working hours
608 increased participants' anxiety by 0.06 points (95% CI 0.00 to 0.11) and burnout by 0.20 points (95% CI
609 0.09 to 0.32). The coefficients for depression and loneliness also suggest a negative effect, but they were
610 smaller and not statistically significant. While resilience levels increased by 0.21 points (95% CI 0.02 to

611 0.41) with the impact of COVID-19, resilience-building activities were affected negatively by -0.13 points
612 (95% CI -0.20 to -0.06). We observed a similar but stronger effect on participants' ability to work during
613 the pandemic for all mental health outcomes and resilience-building activities, but not for resilience
614 levels, ranging from 0.17 points (95% CI 0.06 to 0.28) for loneliness, to 0.95 points (95% CI 0.61 to 1.29)
615 for burnout. Age and gender were significantly associated with both COVID-19 related variables, but the
616 effect was small (odds ratios close to 1). A history of using a chat app for mental health significantly
617 increased resilience-building activities by 0.28 points (95% CI 0.10 to 0.41). Participants who had
618 previously accessed Internet resources for mental health saw significantly greater anxiety (0.17 points,
619 95% CI 0.04 to 0.31) and depression (0.20 points, 95% CI 0.05 to 0.35).

620 While there were differences in outcomes between control and treatment groups, only higher anxiety
621 and loneliness were significantly associated with the treatment group. As expected from the DiD, effect
622 size, and reliable change analyses, there was a significant relationship between study period and all
623 mental wellbeing and resilience outcomes with strong gains post-intervention.

624 For the treatment group, Table 7 shows that the mood score was positively correlated with all mental
625 health outcomes and resilience levels but not with resilience-building activities. A 1-point increase in
626 mood status was associated with a -0.83 point (95% CI -1.04 to -0.61) decrease in anxiety, -0.71 point
627 (95% CI -0.95 to -0.47) decrease in depression, -1.44 point (95% C -1.85 to -1.03) decrease in burnout, -
628 0.38 point (95% CI -0.60 to -0.17) decrease in loneliness, and 2.37 point (95% CI 1.59 to 3.16) increase in
629 resilience. Specific carelines selected by participants had little impact on mental health or resilience
630 outcomes, except when comparing the stress and COVID-19 carelines to the anxiety careline. The stress
631 careline was associated with significantly greater depression but lower burnout than the anxiety
632 careline; there was no significant relationship between the stress careline and other outcomes. The
633 COVID-19 careline was associated with lower mental wellbeing scores for all 4 assessments and higher

resilience-building activities and resilience levels. However, these relationships were only marginally significant at the 0.1 level of statistical significance for depression and loneliness.

637 **Table 7. Additional determinants of mental health and resilience outcomes for the treatment group.**

	Dependent variables (assessment scores)											
	Depression (PHQ-9)		Burnout (OLBI)		Anxiety (GAD-7)		UCLA LONELINESS		Resilience building activities		Resilience (RS-14)	
Independent variables	**Coefficient (SE)**		**Coefficient (SE)**		**Coefficient (SE)**		**Coefficient (SE)**		**Coefficient (SE)**		**Coefficient (SE)**	
Mood score	-0.71***	(0.12)	-1.44***	(0.21)	-0.83***	(0.11)	-0.38***	(0.11)	0.26	(0.19)	2.37***	(0.40)
Careline: depression †	0.20	(0.52)	-1.34	(0.89)	-0.26	(0.46)	-0.09	(0.46)	-1.18*	(0.70)	-3.17*	(1.67)
Careline: stress †	1.06**	(0.43)	-1.79**	(0.74)	0.62	(0.39)	-0.36	(0.39)	-0.02	(0.63)	-0.86	(1.44)
Careline: relationships †	-0.32	(0.43)	-1.03	(0.75)	0.36	(0.38)	-0.15	(0.38)	-0.14	(0.57)	-1.88	(1.44)
Careline: COVID-19 †	-0.60*	(0.35)	-0.82	(0.60)	-0.21	(0.31)	-0.60*	(0.31)	0.40	(0.56)	0.36	(1.15)

638 P-values: * <= 0.10, ** <= 0.05, *** <= 0.01

639 Reference category: † Careline: anxiety

640 ## Participant experience

641 At the end of the study, 820 participants completed an anonymous online participant experience
642 questionnaire (37% from the control group and 63% from the treatment group). The gender distribution
643 was the same (37% male and 63% female). Respondents were predominantly young (54% in the 18-to-
644 29-year age group). Nurses made up the majority (68%). These respondent characteristics are similar to
645 the attributes of the study population.

646 Fifty percent of participants from the treatment group reported using the app for more than 28 days out
647 of the total study period of 56 days (in line with an average of 24 days of use based on app data). When
648 asked why the app was not used more frequently, the top 3 responses were: "I did not have the time to
649 use the app" (45%); "The app took too long to work with" (30%); and "The app was too difficult to use"
650 (13%). Overall, 91% of users found the app easy to use; 87% found the app content relevant; 92%
651 derived benefits from the app; and 83% felt more resilient to everyday challenges. However, Vitalk users
652 encountered several difficulties: 32% found the trial welcome email confusing; 27% found it difficult to
653 download and install the app; 20% found it difficult to sign into the app; 19% had difficulties entering
654 the trial ID; 25% found getting to the mental health assessments confusing; and 33% noted that the app
655 took too long to work with. The top 3 likes of app use were: interactions with Viki (71%); the exercises
656 (58%); and taking the mental wellbeing and resilience assessments (55%). The top three dislikes or
657 challenges of app use were: exposure to new content every or every other day (36%); the mental health
658 exercises (28%); and completing the mental health and resilience assessments (25%). The net promotor
659 score (NPS) [39], was assessed by asking how likely users were on a scale from 1 (not likely) to 10 (very
660 likely) to recommend Vitalk to others. Sixty-six percent out of 515 respondents rated Vitalk a 9 or 10
661 (promoters), 23% a 7 or 8 (passives), and 11% a 6 or less (detractors). This resulted in a NPS – promoters
662 minus detractors – of 55.

Fifty-two percent of participants from the control group reported using the Internet resources for more than 15 days out of the total study period of 56 days, a finding that is much higher than the one gleaned through web analytics (an average of less than 2 days accessing web resources). When asked why the Internet resources were not used more frequently the top 3 responses were: “I did not have the time to use the resources” (46%); “The resources took too long to work with” (34%); and “The content was too difficult to understand” (13%). Overall, 87% of users found the Internet resources easy to use; 92% found the resource content relevant; 95% derived benefits from the resources; and 87% felt more resilient to everyday challenges. However, Internet resource users encountered several difficulties: 17% found the trial welcome email confusing; 22% found it difficult to access the resources; 11% found it difficult to get to the website; 7% had difficulties entering the trial ID; 11% found getting to the mental health assessments confusing; and 43% noted that the resources took too long to work with. The top three likes of web resource use were: taking the mental wellbeing and resilience assessments (61%); reading new content (55%); and the exercises (52%). The top three dislikes or challenges of resource use were: reading the website content (44%); finding new content by following links (36%); and doing mental health exercise (28%). Twenty-three percent found completing the mental health and resilience assessments challenging. When asked how likely users were on a scale from 1 (not likely) to 10 (very likely) to recommend the Internet resources to others, 64% out of 305 respondents rated Internet resources a 9 or 10 (promoters), 23% a 7 or 8 (passives), and 13% a 6 or less (detractors). The resulting NPS of 51 is similar to the treatment group.

Discussion

Based on our literature review, this study is the first RCT of a mental health chat app for health workers combining multiple mental health and resilience outcomes. It was the first to include a measurement of resilience and to explore the potential of a mental health chat app to stimulate resilience-building

686 activities among its users. The RCT took place in a low-resource setting in Southern Africa with very
687 limited access to mental health counseling and therapy. The challenges that Malawi faces meeting
688 population health needs have been further exacerbated by the COVID-19 pandemic.

689 The strengths of our RCT were the recruitment of more than 1,500 health professionals, and the
690 management team's continuous encouragement of participants over the 8-week study period.
691 Participants were successfully randomized into the 2 study arms while maintaining single blinding of the
692 research team. The overall enrollment rate was 94%, with group-specific retention rates of 28%
693 (treatment group) and 40% (control group) (see CONSORT diagram in Fig 3). Compared to other much
694 smaller RCTs of iCBT, retention rates were high due to weekly reminders through WhatsApp and email
695 messages, as well as instant support of users encountering issues during registration, sign in or app use.
696 A trial of adolescents with self-identified anxiety concern (with fewer than 50 participants per group)
697 reported 8-week retention rates of 28% (treatment group) and 58% (control group) [40]. An RCT of iCBT
698 of participants with depression and anxiety symptoms referred to care (substantially different from our
699 study population) reported much higher 8-week retention rates of 82% (treatment group) and 72%
700 (control group) [41].

701 The results from the DiD estimators obtained from mixed linear effect modeling support our working
702 hypothesis that use of a chatbot such as Vitalk can effectively improve mental health and resilience
703 outcomes among health professionals in Malawi. The coefficient for the interaction terms between
704 study group and assessment period varied between -0.68 for depression, -0.58 for burnout, -0.44 for
705 anxiety, 1.22 for resilience building activities, and 1.47 for resilience levels, indicating that active
706 engagement with the app was more effective in improving mental wellbeing (except loneliness) and
707 increasing health worker resilience compared to the passive offering of Internet resources. We are
708 assured that Vitalk app use causes greater improvements in mental wellbeing and resilience than the
709 alternative due to the experimental study design and the size of the DiD estimators and confidence

710 intervals (most do not include zero or absence of a treatment effect). It should be noted that such a
711 strong treatment effect was observed despite considerable technical issues faced by Vitalk users. These
712 challenges likely moderated the effect size as indicated by the disproportionate attrition in the
713 treatment group and the feedback from the anonymous user experience survey, suggesting that this
714 was a very realistic trial of a constantly evolving innovative mobile technology.

715 Based on DiD estimators, Vitalk use did not appear to be associated with changes in loneliness. Possible
716 reasons include: a small effective sample for the treatment group (n = 70); a coarse 3-question
717 instrument that may not be sufficiently sensitive for measuring the effect of Vitalk's relationship
718 careline; or app use in general. Although we explored behavioral theory about loneliness, we could not
719 identify any studies that examined whether iCBT or chatbot use influenced social isolation.

720 Our findings add substantial evidence to existing research. Unlike other published studies, we tested the
721 effectiveness of Vitalk by implementing a large RCT in a LMIC context where access to mental wellbeing
722 and resilience support is extremely limited. Compared to our sample sizes of 200 participants or more
723 per study arm for most of our outcome measures, the other existing RCTs tend to suffer from very small
724 samples of about 30 individuals or fewer. Our RCT is also the first of its kind to be applied to the health
725 workforce, which during normal times is at very high risk for work-related stress and burnout, but even
726 more so during the COVID-19 pandemic. In addition to mental wellbeing outcomes, our study not only
727 measured levels of resilience but also resilience-building activities by participants. We were not able to
728 identify studies that looked at resilience as an outcome to compare them to our findings.

729 While both study groups experienced improvements in mental wellbeing and resilience between
730 baseline and endline, the effect size was consistently larger for the treatment than for the control group,
731 further supporting our working hypothesis that Vitalk is more effective in improving mental health and
732 resilience outcomes than passive Internet resources. While we were not able to compare our DiD results

733 to the published literature, we can compare the effect size, which is commonly reported for RCTs. Our
734 findings suggest an approximately medium effect size, varying between d = -0.36 to -0.52 for mental
735 health, and d = 0.42 to 0.78 for resilience, which is similar to reported literature on the effectiveness of
736 c-CBT of depression and anxiety based on 24 RCTs with sample sizes between 19 and 257 adolescents
737 and young adults with symptoms of anxiety or depression [11]. However, it is smaller than an average
738 effect size of g = 0.77 (95 % CI 0.59–0.95) reported from a meta-analysis of 49 RCTs [9]. An RCT (n = 56)
739 on the use of Woebot (a chatbot) to deliver CBT to young adults with depression and anxiety reported a
740 significant reduction of both outcomes with an effect size of d = 0.44 and 0.37, respectively [42]. An RCT
741 of the use of the Vivibot chatbot (n = 45) by young people with cancer found a significant interaction
742 effect between time and study group of -0.41 for reducing anxiety, but no effect on depression [43]. A
743 review of 12 papers cited weak evidence of the ability of chatbots to reduce depression, with an effect
744 size of -0.55 (95% CI -0.87 to -0.23), which is very similar to our result, but reported outcomes for
745 anxiety contradicting our findings [19]. Effect size may be influenced by the characteristics of the study
746 population. While our study targeted healthy healthcare workers, other studies enrolled participants
747 with predetermined mental health issues or underlying illnesses. The fact that healthcare workers have
748 self-selected to cope with stressful situations compared to participants from the general population, or
749 those clinically diagnosed with symptoms, may explain some of the discrepancies in effect size.

750 Nonexperimental studies showed a greater effect size than RCTs. An earlier effectiveness study of self-
751 selected Vitalk users from the general population reported a large effect size of d = 0.81 and greater for
752 mental health outcomes [16]. Another uncontrolled trial of clients with severe symptoms of depression
753 and anxiety reported much higher scores for depression and anxiety (based on PHQ-9 and GAD-7
754 assessments) than our study, as well as effect sizes ranging from d = 0.61 to 0.83 [44].

755 The LMIC setting of our RCT may account for a smaller effect size than anticipated compared to results
756 from studies in middle- and high-income countries. Limited and expensive access to mobile phones in

757 Malawi provides less opportunity for self-help and other innovative app use among the population,
758 which may have constrained participants' ability to benefit from the full potential of a mental health
759 chatbot. This is supported by the fact that a high proportion of participants reported that they did not
760 have the time to use the app or Internet resources (1 in 5), or found that it took too long to work with
761 them (1 in 3). This could also imply that participants tried to conserve precious airtime. Large differences
762 in mean scores between our study and others suggest that culture may also play an important role in
763 shaping how participants respond to mental health and resilience questionnaires [45, 46]. For example,
764 our RCT in Malawi reported mean anxiety scores (4.5 at baseline and 2.9 at endline) and depression
765 scores (4.0/2.9) that were less than half of those found in Brazil (pre-/post-app use means of 12.2/8.1
766 for anxiety, and 15.9/10.4 for depression) [16].

767 Many of the improvements in mental health and resiliency scores documented in our study are of
768 statistical and clinical relevance and are welcome outcomes. Changes in risk levels were based on RCIs of
769 4 for the GAD-7 (anxiety) and PHQ-9 (depression) calculated from our sample. These are similar to or
770 lower than those reported by other studies (e.g., an RCI of 6 for PHQ-9 and an RCI of 4 for GAD-7 [44],
771 and an RCI of 6 for GAD-7 [47]). Based on our RCIs, we estimate that about one- to two-thirds of the
772 improvements in both study groups were reliable (see Fig 9). Applying the RCI, we saw reliable
773 improvements for depression of 14% (control group) and 17% (treatment group); and improvements in
774 anxiety of about 25% (both groups). This is considerably lower than the reliable change in clinical
775 caseness reported by other studies (e.g., 52% for anxiety and 45% for depression [16] and between 58%
776 and 70% across assessments [44]). These discrepancies are likely due to differences in study populations
777 (randomly selected health workers versus self-selected general population), with much lower mean
778 scores for our participants, and a small proportion classified at moderate or severe risk levels of about
779 13% at baseline and 8% at endline for depression and anxiety, respectively. A small proportion (0% to
780 7%) of participants saw a reliable worsening of mental health and resilience scores; the effect was

781 consistently higher in the control group. This decline is in line with about 2% to 5% reported by Daley et
782 al. [16].

783 While controlling for trial group affiliation and study period, our analysis explored which user or
784 engagement characteristics most significantly explained the outcomes of interest. Although these
785 factors may vary by country context, they must be considered when planning mental health support
786 strategies and helping health workers increase their resilience to meet everyday challenges.
787 Engagement with the app or the website, measured as z-scores of time spent, was strongly correlated
788 with reduced anxiety, depression, burnout, and loneliness. Gan et al. [48] also found a significant
789 association between engagement in digital mental health interventions and mental health
790 improvements. A systematic review by Molly and Anderson [49] identified several studies that showed
791 statistically significant associations between engagement and clinical improvement. Our data did not
792 indicate that time was a determinant of resilience or resilience-building activities.

793 Our study identified a need for user interface and app processes to be designed for people with less
794 experience using smartphones and online resources and with a higher risk of adverse mental health
795 outcomes. In our study, we found that older and female participants tended to have greater anxiety,
796 depression, loneliness (age only), and burnout (gender only), as well as less experience with smartphone
797 and Internet use. In a study of the health workforce in Jordan, age and gender were associated with
798 worse mental health outcomes for males [50]. A rapid scoping review of occupational stress, burnout,
799 and depression in women in healthcare during the COVID-19 pandemic showed that female health
800 workers were at increased risk [51].

801 Compared to those working in other types of medical and ancillary services, levels of anxiety and
802 depression appear to be highest among our participants working in outpatient care, a service in which
803 healthcare workers may be required to pay special attention to mental health support, but, according to

804 anonymous user feedback, they may also face the greatest time constraints for accessing online mental
805 health solutions. Participants working in secondary and tertiary care institutions were at higher risk for
806 anxiety, depression, and burnout compared to those in primary care, a finding similar to a study in Peru
807 [52]. In our study, we found that public facilities were associated with greater mental health risks than
808 private, not-for-profit facilities. Public facilities may be able to draw upon the strategies used by private
809 facilities to reduce anxiety and depression. However, the public sector in Malawi seems better equipped
810 to build health worker resilience compared to the private sector. Work location (urban versus rural), was
811 not significantly associated with the outcomes (at the 95% level).

812 Our findings that current or past exposure to mental health counseling was associated with greater
813 anxiety, depression and burnout confirmed those of Varma et al. [53]. Our participants saw a negative
814 impact of the COVID-19 pandemic across all outcomes (except for resilience scores), similarly to findings
815 from a study in Jordan [50]. A study of healthcare and allied workers involved in the COVID-19 pandemic
816 from India found high levels of severe and extremely severe anxiety (23.2%) and severe and extremely
817 severe depression (11.4%)[54]. The level of severe anxiety was almost double of that found in our study
818 (which did not focus on workers involved in the pandemic). The levels of severe depression were similar
819 to our findings. Although the Vitalk app demonstrated no effect on loneliness, the pandemic itself had a
820 negative impact on loneliness overall. This could be expected based on the results of a meta-analysis
821 reporting a small but significant effect size of $g = 0.27$ of the COVID-19 pandemic on loneliness [55].

822 Our study had several limitations. Due to technical issues with the app, the response rate for the
823 treatment group was substantially lower than for the control group. The attrition rate was greater for
824 the loneliness and resilience-building activity assessments, reducing the power of our statistical analysis.
825 Engagement with the app was an average of 11 minutes per day, including the completion of
826 assessments, and was within a range of 4 to 22 minutes (median) reported by a study of user
827 engagement with mental health apps [56]. The RCT provided participants with airtime to use Vitalk or to

828 peruse the web resources on their mobile phones, allowing us to test a mental wellbeing digital
829 intervention in a resource-poor environment in anticipation of a future rapid increase of affordable
830 cellular data access in sub-Saharan Africa. Our study was not representative of the entire health
831 workforce in Malawi, because the interventions required English language proficiency and smartphone
832 ownership, limiting the study to those with a diploma or at least 2 years of college education. Moreover,
833 participants were predominantly from 2 urban districts within major population centers and only 18%
834 from rural health facilities. Given that the classification of mental health and resilience relied entirely on
835 self-assessments, which can be influenced by sociocultural factors unique to a country or study
836 population, this phenomenon must be taken into account when comparing our findings to those of
837 other studies.

838 As indicated by NPS scores, user satisfaction with Vitalk and web resources was high in the treatment
839 (NPS of 55) and control (NPS of 51) groups. While we did not find NPS benchmarks for chat apps or iCBT,
840 a study by Marler [57] compared the NPS in an RCT of 2 smoking cessation apps and found values
841 ranging between 1 (control app) and 57 (intervention app). A study of community mental health services
842 found an NPS of 19, which varied considerably by respondent characteristics (from 10 for older, to 37 for
843 younger patients) [29]. Nonetheless, given the limitations of validity and reliability of single item
844 measurements such as the NPS [58], the magnitude of the NPS and difference between groups should
845 be viewed with caution. Moreover, apps such as Vitalk are constantly updated and enhanced as the
846 technology matures. During the RCT in Malawi, a newer Vitalk version with advanced artificial
847 intelligence and natural language processing features offering a potentially better user experience was
848 already under development in Brazil.

Conclusions

Given Vitalk's demonstrably positive effect on mental wellbeing and resilience, it has great potential to be an effective tool in the arsenal of interventions to reduce work-related stress and burnout. Based on our completion rate (1 in 3 study participants), there appears to be substantial demand for such an app and willingness to engage with it for at least 10 minutes daily (a length of time that seems adequate for most users to improve mental wellbeing and resilience, in addition to extra time for occasional assessments). User feedback suggests that several factors encourage chatbot use including: natural interaction with the app; frequently updated content and exercises; and carelines adapted to 2 diverse types of users – those needing more general psychosocial support and those facing more pronounced mental health issues. Users indicated that they enjoyed the interaction with the virtual assistant, Viki, and appreciated the assessments on mental wellbeing and resilience as important benchmarks. Making the assessments less tedious and more entertaining is an ongoing challenge for app developers. App developers should weigh the advantages and disadvantages of changing the clinical risk classifications based on standardized mental health and resilience assessments from their original to fixed groupings, even if this makes sense from a user-friendliness perspective.

While apps such as Vitalk can help health professionals cope with daily stressors and improve mental wellbeing, employers need to create less stressful work environments to maximize their effectiveness. Proven interventions to strengthen health systems function and to increase healthcare provider satisfaction include ensuring: essential supplies and equipment; an infrastructure that meets the needs of patients and staff; and adequate and well-trained human resources for health. As Internet access through mobile devices becomes more widespread and affordable in LMICs, employers in the public and private sector should consider providing free app subscriptions to their employees, organizing group sessions for mental wellbeing exercises, and encouraging app use.

872 Similarly to other studies, our RCT confirmed that mobile chat apps are one effective tool for improving
873 mental wellbeing and resilience among its target audience. For the first time, our findings show that
874 apps such as Vitalk can support the mental wellbeing and resilience of health workers to meet the
875 challenges unique to their profession, especially in a resource poor environment where access to mental
876 health services is very limited. Countries employ thousands of health workers, and many need support:
877 as many as 1 in 8 health workers in our RCT needed help with anxiety or depression; 3 in 4 faced some
878 degree of burnout; and 1 in 4 had low resilience. As countries struggle to increase access to mental
879 health professionals, apps such as Vitalk have potential to fill the gaps in support. In the aftermath of
880 the ongoing COVID-19 pandemic, healthcare providers in the public and private sectors should consider
881 offering access to mental wellbeing apps to all their employees to cope with everyday challenges and to
882 prepare for future public health emergencies.

883 Acknowledgments

884 The authors would like to thank the management and health professionals from health centers and
885 hospitals in Malawi for their support and participation in this study. We gratefully acknowledge
886 Elizabeth Leitman (Chemonics International) for editing the manuscript.

887 Disclaimer

888 The views and opinions expressed in this paper are those of the authors and not necessarily the views
889 and opinions of the United States Agency for International Development.

Funding

Financial support for this RCT was provided by the United States Agency for International Development under the Cooperative Agreement No. AID-OAA-A-15-00046. The funders had no involvement in the analysis or in the decision to submit the paper for publication.

Role of the funding source

The study was funded by the U.S. Agency for International Development (USAID) through the Human Resources for Health 2030 Project, led by Chemonics International in partnership with University Research Company (URC) and other partners. USAID had no role in the data collection, data analysis, data interpretation, or writing of the paper. All authors had full access to all de-identified study data and had final responsibility for the decision to submit for publication.

Contributions

EK and TL designed the study and the resilience-building questionnaire, performed data cleaning and analysis, and drafted the manuscript. TL designed the study, led the research team in Malawi, managed the participant workshops, and provided feedback on the manuscript. WJ provided detailed input to the protocol and feedback on the manuscript. KG led the study management team, coordinated participant communications and responses to support requests, curated the data, and provided feedback on the manuscript. IH provided input to the research protocol. DR managed the communication with and data reporting by Vitalk. DK, LM, EJ, and AZ provided feedback on the protocol and manuscript and facilitated participant workshops in Malawi. MN was the Vitalk head of product. RM worked on the data and app/web analytics. YK was the Vitalk project lead. MK oversaw the Vitalk design and development.

References

1. Sovold LE, Naslund JA, Kousoulis AA, Saxena S, Qoronfleh MW, Grobler C, et al. Prioritizing the Mental Health and Well-Being of Healthcare Workers: An Urgent Global Public Health Priority. Front Public Health. 2021;9:679397.
2. Muller AE, Hafstad EV, Himmels JPW, Smedslund G, Flottorp S, Stensland SO, et al. The mental health impact of the covid-19 pandemic on healthcare workers, and interventions to help them: A rapid systematic review. Psychiatry Res. 2020;293:113441.
3. Nochaiwong S, Ruengorn C, Thavorn K, Hutton B, Awiphan R, Phosuya C, et al. Global prevalence of mental health issues among the general population during the coronavirus disease-2019 pandemic: a systematic review and meta-analysis. Sci Rep. 2021;11(1):10173.
4. Ofori AA, Osarfo J, Agbeno EK, Manu DO, Amoah E. Psychological impact of COVID-19 on health workers in Ghana: A multicentre, cross-sectional study. SAGE Open Med. 2021;9:20503121211000919.
5. Mulatu HA, Tesfaye M, Woldeyes E, Bayisa T, Fisseha H, Kassu RA. The prevalence of common mental disorders among healthcare professionals during the COVID-19 pandemic at a tertiary Hospital in Addis Ababa, Ethiopia. J Affect Disord Rep. 2021;6:100246.
6. Msomi N. SA healthcare workers been experiencing burnout long before Covid-19 pandemic: News24, health 24; 2021 [Available from: https://www.news24.com/health24/mental-health/mental-health-in-sa/sa-healthcare-workers-been-experiencing-burnout-long-before-covid-19-pandemic-20210318-2.
7. Chorwe-Sungani G. Assessing COVID-19-related anxiety and functional impairment amongst nurses in Malawi. Afr J Prim Health Care Fam Med. 2021;13(1):e1-e6.
8. Chisholm D, Sweeny K, Sheehan P, Rasmussen B, Smit F, Cuijpers P, et al. Scaling-up treatment of depression and anxiety: a global return on investment analysis. Lancet Psychiatry. 2016;3(5):415-24.
9. Grist R, Cavanagh K. Computerised Cognitive Behavioural Therapy for Common Mental Health Disorders, What Works, for Whom Under What Circumstances? A Systematic Review and Meta-analysis. Journal of Contemporary Psychotherapy. 2013;43:243-51.
10. Andrews G, Cuijpers P, Craske MG, McEvoy P, Titov N. Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis. PLoS One. 2010;5(10):e13196.
11. Christ C, Schouten MJ, Blankers M, van Schaik DJ, Beekman AT, Wisman MA, et al. Internet and Computer-Based Cognitive Behavioral Therapy for Anxiety and Depression in Adolescents and Young Adults: Systematic Review and Meta-Analysis. J Med Internet Res. 2020;22(9):e17831.
12. Watts S, Mackenzie A, Thomas C, Griskaitis A, Mewton L, Williams A, et al. CBT for depression: a pilot RCT comparing mobile phone vs. computer. BMC Psychiatry. 2013;13:49.
13. Kristjansdottir OB, Fors EA, Eide E, Finset A, Stensrud TL, van Dulmen S, et al. A smartphone-based intervention with diaries and therapist-feedback to reduce catastrophizing and increase functioning in women with chronic widespread pain: randomized controlled trial. J Med Internet Res. 2013;15(1):e5.
14. Dagoo J, Asplund RP, Bsenko HA, Hjerling S, Holmberg A, Westh S, et al. Cognitive behavior therapy versus interpersonal psychotherapy for social anxiety disorder delivered via smartphone and computer: a randomized controlled trial. J Anxiety Disord. 2014;28(4):410-7.
15. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental Health Smartphone Apps: Review and Evidence-Based Recommendations for Future Developments. JMIR Ment Health. 2016;3(1):e7.

16. Daley K, Hungerbuehler I, Cavanagh K, Claro HG, Swinton PA, Kapps M. Preliminary Evaluation of the Engagement and Effectiveness of a Mental Health Chatbot. Front Digit Health. 2020;2:576361.
17. Mental health atlas 2017. World Health Organization; 2018.
18. Kauye F. Management of mental health services in Malawi. Int Psychiatry. 2008;5(2):29-30.
19. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. J Med Internet Res. 2020;22(7):e16021.
20. Dallal GE. Randomization plan generator 2020 [Available from: http://www.jerrydallal.com/random/randomize.htm.
21. Research Randomizer: Social Psychology Network; 2021 [Available from: https://www.randomizer.org/.
22. Seligman MEP, editor Building Human Strength: Psychology's Forgotten Mission. 2006.
23. Beck JS, Beck AT. Cognitive Behavior Therapy, Second Edition: Basics and Beyond: Guilford Publications; 2011.
24. Yoo B-k, Rayz JT. Understanding Emojis for Sentiment Analysis. The International FLAIRS Conference Proceedings. 2021.
25. Looking after our mental health: Mental Health Foundation (MHF); 2021 [Available from: https://www.mentalhealth.org.uk/publications/how-to-mental-health.
26. Mental well-being: resources for the public: World Health Organization; 2021 [Available from: https://www.who.int/news-room/feature-stories/mental-well-being-resources-for-the-public.
27. #HealthyAtHome - Mental health: World Health Organization; 2020 [Available from: https://www.who.int/campaigns/connecting-the-world-to-combat-coronavirus/healthyathome/healthyathome---mental-health.
28. Doing what matters in times of stress: an illustrated guide. Geneva: World Health Organization; 2020.
29. Wilberforce M, Poll S, Langham H, Worden A, Challis D. Measuring the patient experience in community mental health services for older people: A study of the Net Promoter Score using the Friends and Family Test in England. Int J Geriatr Psychiatry. 2019;34(1):31-7.
30. Demerouti E, Bakker AB. The Oldenburg Burnout Inventory: A good alternative to measure burnout and engagement. Halbesleben JRB, editor. Hauppauge, NY: Nova Science Pub Inc.; 2008.
31. Kaggwa MM, Kajjimu J, Sserunkuma J, Najjuka SM, Atim LM, Olum R, et al. Prevalence of burnout among university students in low- and middle-income countries: A systematic review and meta-analysis. PLoS One. 2021;16(8):e0256402.
32. Mughal AY, Devadas J, Ardman E, Levis B, Go VF, Gaynes BN. A systematic review of validated screening tools for anxiety disorders and PTSD in low to middle income countries. BMC Psychiatry. 2020;20(1):338.
33. Steptoe A, Shankar A, Demakakos P, Wardle J. Social isolation, loneliness, and all-cause mortality in older men and women. Proc Natl Acad Sci U S A. 2013;110(15):5797-801.
34. Hughes ME, Waite LJ, Hawkley LC, Cacioppo JT. A Short Scale for Measuring Loneliness in Large Surveys: Results From Two Population-Based Studies. Res Aging. 2004;26(6):655-72.
35. Building Health Worker Resilience: A Toolkit to Protect Against Burnout on the Front Lines. New Brunswick, New Jersey, USA: Center for Health Worker Innovation; 2021.
36. Wagnild G. The Resilience Scale user's guide for the US English version of the Resilience Scale and the 14-item Resilience Scale (RS–14). Worden, MT: Resilience Center; 2009.
37. Wagnild G. A review of the Resilience Scale. J Nurs Meas. 2009;17(2):105-13.

1003 38. Siriwardhana C, Abas M, Siribaddana S, Sumathipala A, Stewart R. Dynamics of resilience in
1004 forced migration: a 1-year follow-up study of longitudinal associations with mental health in a
1005 conflict-affected, ethnic Muslim population. BMJ Open. 2015;5(2):e006000.
1006 39. Reichheld FF. The one number you need to grow. Harv Bus Rev. 2003;81(12):46-54, 124.
1007 40. O'Connor K, Bagnell A, McGrath P, Wozney L, Radomski A, Rosychuk RJ, et al. An Internet-Based
1008 Cognitive Behavioral Program for Adolescents With Anxiety: Pilot Randomized Controlled Trial.
1009 JMIR Ment Health. 2020;7(7):e13356.
1010 41. Richards D, Enrique A, Eilert N, Franklin M, Palacios J, Duffy D, et al. A pragmatic randomized
1011 waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for
1012 depression and anxiety. NPJ Digit Med. 2020;3:85.
1013 42. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With
1014 Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot):
1015 A Randomized Controlled Trial. JMIR Ment Health. 2017;4(2):e19.
1016 43. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the Chatbot "Vivibot" to
1017 Deliver Positive Psychology Skills and Promote Well-Being Among Young People After Cancer
1018 Treatment: Randomized Controlled Feasibility Trial. JMIR Mhealth Uhealth. 2019;7(10):e15018.
1019 44. Duffy D, Enrique A, Connell S, Connolly C, Richards D. Internet-Delivered Cognitive Behavior
1020 Therapy as a Prequel to Face-To-Face Therapy for Depression and Anxiety: A Naturalistic
1021 Observation. Front Psychiatry. 2019;10:902.
1022 45. Parkerson HA, Thibodeau MA, Brandt CP, Zvolensky MJ, Asmundson GJ. Cultural-based biases of
1023 the GAD-7. J Anxiety Disord. 2015;31:38-42.
1024 46. Harry ML, Coley RY, Waring SC, Simon GE. Evaluating the Cross-Cultural Measurement
1025 Invariance of the PHQ-9 between American Indian/Alaska Native Adults and Diverse Racial and
1026 Ethnic Groups. J Affect Disord Rep. 2021;4.
1027 47. Bischoff T, Anderson SR, Heafner J, Tambling R. Establishment of a reliable change index for the
1028 GAD-7. Psychology, community & health. 2020;8:176-87.
1029 48. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M. Effect of Engagement With Digital
1030 Interventions on Mental Health Outcomes: A Systematic Review and Meta-Analysis. Front Digit
1031 Health. 2021;3:764079.
1032 49. Molloy A, Anderson PL. Engagement with mobile health interventions for depression: A
1033 systematic review. Internet Interv. 2021;26:100454.
1034 50. Alnazly E, Khraisat OM, Al-Bashaireh AM, Bryant CL. Anxiety, depression, stress, fear and social
1035 support during COVID-19 pandemic among Jordanian healthcare workers. PLoS One.
1036 2021;16(3):e0247679.
1037 51. Sriharan A, Ratnapalan S, Tricco AC, Lupea D, Ayala AP, Pang H, et al. Occupational Stress,
1038 Burnout, and Depression in Women in Healthcare During COVID-19 Pandemic: Rapid Scoping
1039 Review. Front Glob Womens Health. 2020;1:596690.
1040 52. Quispe-Sancho A, Chambi-Macedo KL, Laurel-Vargas V, Huamani-Merma E, Cuzcano-Gonzales
1041 KV, Huaita-Rocha MA, et al. Depression, Anxiety, and Stress in Health Professionals Working
1042 During the COVID-19 Pandemic in Peru: An Analytical Cross-Sectional Study. ELECTRON J GEN
1043 MED. 2021;18(6).
1044 53. Varma P, Junge M, Meaklim H, Jackson ML. Younger people are more vulnerable to stress,
1045 anxiety and depression during COVID-19 pandemic: A global cross-sectional survey. Prog
1046 Neuropsychopharmacol Biol Psychiatry. 2021;109:110236.
1047 54. Shekhar S, Ahmad S, Ranjan A, Pandey S, Ayub A, Kumar P. Assessment of depression, anxiety
1048 and stress experienced by health care and allied workers involved in SARS-CoV2 pandemic. J
1049 Family Med Prim Care. 2022;11(2):466-71.

1050 55. Ernst M, Niederer D, Werner AM, Czaja SJ, Mikton C, Ong AD, et al. Loneliness before and during
1051 the COVID-19 pandemic: A systematic review with meta-analysis. Am Psychol. 2022;77(5):660-
1052 77.
1053 56. Baumel A, Muench F, Edan S, Kane JM. Objective User Engagement With Mental Health Apps:
1054 Systematic Search and Panel-Based Usage Analysis. J Med Internet Res. 2019;21(9):e14567.
1055 57. Marler JD, Fujii CA, Utley MT, Balbierz DJ, Galanko JA, Utley DS. Outcomes of a Comprehensive
1056 Mobile Smoking Cessation Program With Nicotine Replacement Therapy in Adult Smokers: Pilot
1057 Randomized Controlled Trial. JMIR Mhealth Uhealth. 2022;10(11):e41658.
1058 58. Adams C, Walpola R, Schembri AM, Harrison R. The ultimate question? Evaluating the use of Net
1059 Promoter Score in healthcare: A systematic review. Health Expect. 2022;25(5):2328-39.

1060