University of Bath

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

User-centered Development of STOP (Successful Treatment for Paranoia): Material Development and Usability Testing for a Digital Therapeutic for Paranoia

Che-Wei Hsu[1,8]*, Daniel Stahl[2], Elias Mouchlianitis[1], Emmanuelle Peters[3,7], George Vamvakas[2], Jeroen Keppens[4], Miles Watson[1], Nora Schmidt[1], Pamela Jacobsen[5], Philip McGuire[1], Sukhwinder Shergill[1], Thomas Kabir[6], Tia Hirani[1], Ziyang Yang[1], and Jenny Yiend[1]*.

[1]Psychosis Studies, King's College London

[2] Bio Statistics and Health Informatics, King's College London

[3] Department of Psychology, King's College London

[4] Department of Informatics, King's College London

[5] University of Bath

[6] McPin Charity Foundation

[7] South London and Maudsley NHS Foundation Trust (SLaM)

[8] Department of Psychological Medicine, University of Otago

*Corresponding author: Professor Jenny Yiend, PO 63, Institute of Psychiatry, King's College London, De Crespigny Park, London, SE5 8AF. Tel 0207848 5119. Email: jenny.yiend@kcl.ac.uk. Phone: +44 (020)78485119

Abstract

Background

Paranoia is a highly debilitating mental health condition. One novel intervention for paranoia is cognitive bias modification for paranoia (CBM-pa). CBM-pa comes from a class of interventions that focus on manipulating interpretation bias. Here, we aimed to develop and evaluate new therapy content for CBM-pa for later use in a self-administered digital therapeutic for paranoia called STOP ('Successful Treatment of Paranoia').

Objectives

The present study had the following objectives:

1) take a user-centered approach with input from living experts, clinicians, and academics, to create and evaluate paranoia-relevant item content to be used in STOP

2) engage with living experts and the design team from a digital healthcare solutions company to co-create and pilot test the STOP mobile app prototype.

Methods

Content Development. We invited 18 people with living/lived experience of paranoia to create text exemplars of personal, everyday emotionally ambiguous scenarios that could provoke paranoid thoughts. Researchers then adapted 240 suitable exemplars into corresponding intervention items in the format commonly used for CBM training and created 240 control items for the purpose of testing STOP. Each item included newly developed, visually enriching graphics content to increase the engagement and realism of the basic text scenarios. All items were then evaluated by living experts (n = 8) and clinicians (n = 7) for paranoia severity, readability, and by the research team for item length. Items were evenly distributed into six, 40-item sessions based on these evaluations.

Usability Testing. Finalized items were presented in the STOP mobile app, which was co-designed with a digital healthcare solutions company, living/lived experts, and the

academic team, and user acceptance was evaluated across two pilot tests involving living/lived experts.

Results

Content Development. All materials reached predefined acceptable thresholds on all rating criteria: *paranoia severity (intervention items:* $\geq$ 1; control items: $\leq$ 1*, readability:* $\geq$ 3*, and length of the scenarios*, and there was no systematic difference between intervention and control group materials overall or between individual sessions within each group. For item graphics, we also found no systematic differences in users' ratings of complexity ($p = .68$) attractiveness ($p = .15$), and interest ($p = .14$)*,* between intervention and control group materials.

Usability Testing. User acceptance testing of the mobile app found the application easy to use and navigate, interactive, and helpful.

Conclusions. Material development for any new digital therapeutic requires an iterative and rigorous process of testing involving multiple contributing groups. Appropriate user-centered development can create user-friendly mobile health applications, which may improve face validity, and have a greater chance to be engaging and acceptable to the target end users.

Introduction

Psychosis is one of the most disabling mental health conditions presenting with significant distress, suicidal ideation, impaired social and occupational functioning, and physical ill-health [1,2]. Paranoia and associated delusions are common symptoms of psychosis, are associated with more distress than other types of delusion [3], are most likely to be acted upon [4], and represent a strong predictor of hospitalization [5]. Over one-third of all UK psychiatric patients suffer from paranoia, which also presents in a range of other psychopathologies, including depression [6], bipolar disorder [7], posttraumatic stress disorder [8], anxiety [9], as well as schizophrenia [10].

The National Institute for Health and Care Excellence recommended Cognitive Behavioural Therapy (CBT) for treating psychosis. CBT, however, is received by only 1 in 10 of those who could benefit and has shown only moderate effect sizes for the treatment of delusions [11,12], although effect sizes are higher for those studies targeting delusions specifically, as opposed to generic CBT [13,14]. Unfortunately, a significant proportion of patients suffering from paranoia continue to experience distressing symptoms following psychological treatment [15,16]. Consequently, there is a need for novel, highly accessible, and low-cost interventions for paranoia, either as standalone treatments or as adjuncts to boost existing therapies. Cognitive bias modification (CBM) is a class of intervention that may address these needs.

Cognitive bias modification

The class of CBM interventions works on the premise that cognitive bias is a putative causal factor of various mental health concerns [17,18,19,20,21]. One form of cognitive bias is interpretation bias, which is the tendency for individuals to think about a situation in a negatively skewed direction. The same situation could also be interpreted in a benign or positive direction. Repeated negatively biased interpretations are thought to contribute to the

formation and maintenance of psychological symptoms and increase distress [22]. Across many studies, researchers have found evidence of interpretation bias among anxiety [19], depression [20], and social phobia [21], with some work on interpretation bias in paranoia [22,23,24,25,26].

CBM is a class of targeted treatment that focuses on manipulating naturally occurring interpretation bias in a more helpful direction, with findings from many studies demonstrating the positive efficacy of CBM with various psychiatric disorders, including anxiety, affective disorders, and substance addictions [27,28,29]. There are several benefits to CBM. First, CBM can be self-administered and disseminated over numerous settings [30], reducing the need for mental health professionals. Next, CBM has the potential to benefit patients whose symptoms may influence their trust in a therapist [31]. Third, CBM can be delivered on a digital platform, which means that it is highly accessible at a low cost [32,33].

Despite these benefits and positive efficacy of CBM with various mental health concerns, there is a dearth of studies on CBM that addresses psychosis, with only some preliminary evidence of the feasibility and implications of this approach. For example, Steel et al. [34] demonstrated the effects of *CBM on anxiety* in individuals diagnosed with schizophrenia. Results from that study showed that a subgroup of participants exhibited positive changes in interpretation bias. Turner et al.'s [35] case study on patients who experienced *social anxiety following a psychotic episode* demonstrated similar positive changes in interpretation bias. In a feasibility study, Yiend et al. [36] directly examined the effects of CBM in patients with paranoia, using an intervention called CBM-pa. In that study, 63 participants with clinically significant persecutory or paranoid symptoms were randomly assigned to either the CBM-pa group (n = 32) or the control group (n = 31). Participants in the CBM-pa group were presented with 40 short passages over six weekly sessions on a computer using a software called E-prime. Users were invited to complete the final word of

each passage, which contained missing letters. Once completed, the word resolved the ambiguity of the passage in a benign, non-paranoid manner. A follow-up yes/no question reinforced the benign interpretation of the passage (see Figure 2 for an example). Sessions were self-directed as users completed each word task independently on the computer. The control group received the same number of sessions over six weeks that included items of general knowledge and facts, and everyday activities. Results showed that relative to the control group, participants in the CBM-pa group showed larger reductions in negative interpretation bias and paranoid symptoms.

Each passage of CBM-pa depicted an emotionally ambiguous scenario, all of which were developed with a user-centered approach, by inviting living experts and experienced clinicians to review all training materials to ensure the clinical relevance of the items to paranoia.

User-Centered Development

Researchers have shown that people experiencing psychosis can benefit from digital therapeutics, but despite the wide availability of digital therapeutics on the App market, many have insufficient evidence-based data to support their efficacy, design, and development [37]. It is important to take a user-centered development approach to design a user-friendly, engaging, and self-managing digital therapeutic for psychosis [38,39] by involving multiple collaborators, including service users, researchers, and the design team. This approach is known to increase the adoption of the App by end users [39] and improve App design and content [40,41]. Self-administered mobile health applications without quality evidence-based data to support their use may decrease the usability and effectiveness of the treatment [42]. This is important for both App design as well as the intervention content. Researchers have demonstrated that biases are stronger when the information being processed is more directly relevant to the disorder at hand [43,44]. Yiend et al. [36] used content-specific training

materials for paranoia to capture and modify paranoia interpretation bias commonly experienced by patients with *paranoid* symptoms. Content materials were co-designed with relevant contributors and sessions were presented in rank order of increasing severity of items using Freeman et al.'s [45] hierarchy of paranoia as a guide. The training items covered six categories relevant to paranoia: social/interpersonal threat, delusions of reference/magical thinking, threat of persecution/spying, general suspiciousness/distrust, medical/paramedical/health care threat, and physical harm.

The Current Study

Building on from Yiend et al. [36] and following a user-centered development approach, here, we aimed to develop CBM-pa into a 12-session mobile App therapeutic called STOP ('Successful Treatment of Paranoia'). As a part of a clinical trial, we tested STOP's efficacy against a control group. STOP included the original item content from the CBM-pa feasibility study and newly developed items for six additional training sessions (details of content development for the six training sessions from the CBM-pa feasibility study will be reported separately). In the present paper, we reported the detailed development process of STOP, which had the following objectives:

1) take a user-centered approach with input from living/lived experts, clinicians, and academics, to create and evaluate paranoia-relevant item content to be used in STOP

2) engage with living/lived experts and the design team from Avegen to co-create and pilot test the STOP mobile app prototype. Avegen is a digital healthcare company speicalizing in developing innovative healthcare technologies [46].

The methodology of the STOP development process involved:

1) four stages for **objective 1:** text creation, text evaluation, graphics development, and graphics evaluation, and

2) one stage for **objective 2**: STOP mobile phone app usability testing.

Objective 1 was intended to ensure clinical relevance, content specificity to paranoia, face validity of the training materials, and user acceptability for STOP. Objective 2 provided data on living/lived experts' perspectives on the functionality, interface, and acceptability of the prototype STOP app to reveal areas of strength and those that needed improvement.

STOP Content Design (Test and Images)

The process of material development and testing spanned 12 months and involved extensive iterative input from 1) living/lived experts, 2) Clinical Psychologists, and 3) the STOP academic research team (see Supplementary Material for inclusion criteria of each contributor). Input from each contributing group and the numbers varied according to the task required and are given below. Figure 1 shows a schematic of the development process.

Stage I

Living/lived experts provide exemplars of everyday ambiguous life scenarios that typically provoke paranoid thoughts

Researchers adapt exemplars into CBM format

Match item length across a) experimental group (STOP vs control) and b) sessions

Stage II

Rate items based on predefined criteria: *severity and readability*

Re-write and re-rate items that do not meet a predefined threshold

Re-match item length across experimental groups and sessions

Stage III

Rate graphics based on predefined criteria: *complexity, attractiveness,* and *interest*

Stage IV

STOP mobile phone app usability testing: two pilot tests involving expert by experience to provide quantitative ratings on the following features of the mobile app: *Ease of Use, User Interface, Interactive Features, Design and Graphics, Security and Privacy, Errors/ Bugs, 'Help' provision* (pilot 2 only).
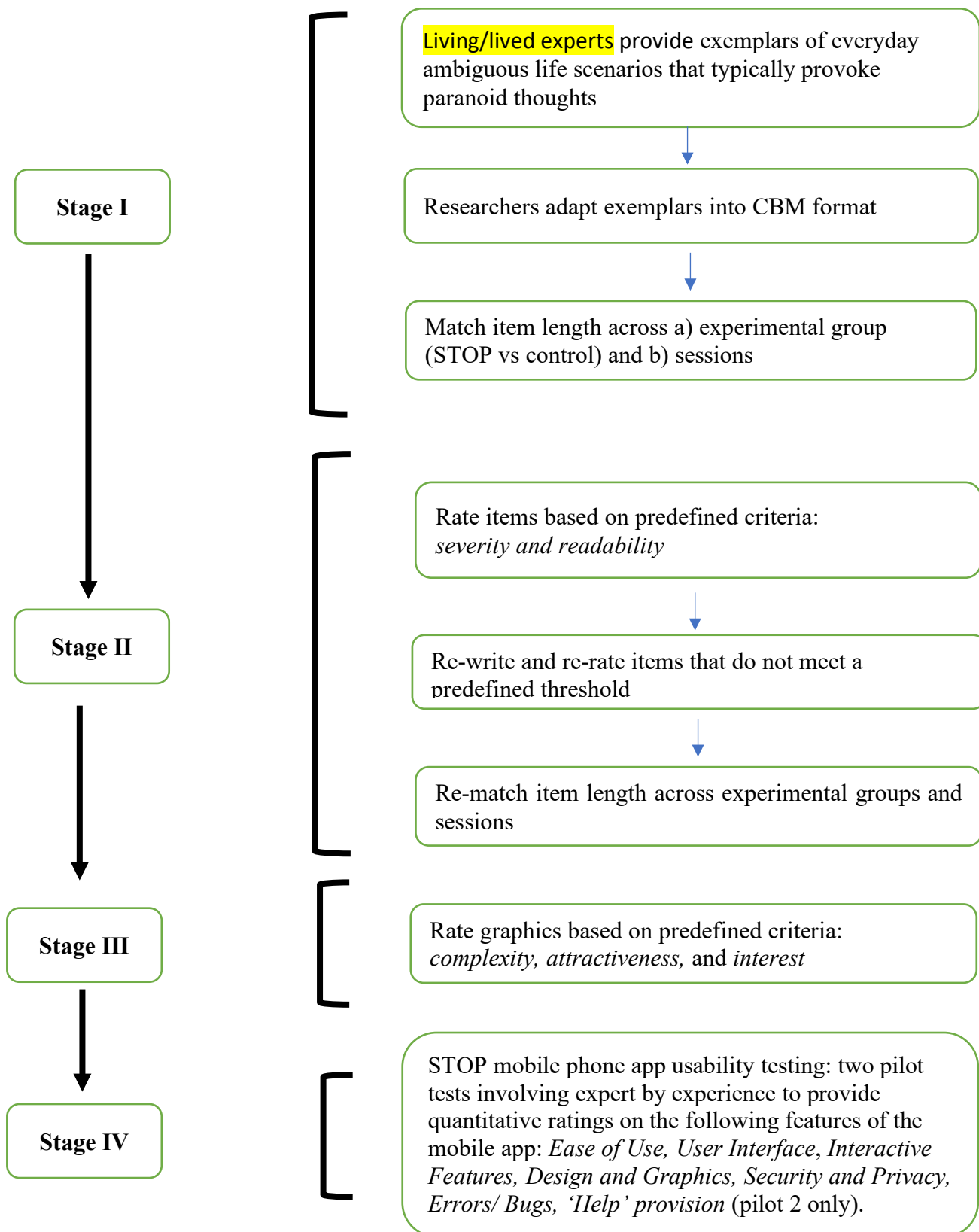
Figure 1. Schematic of STOP materials development process

Stage I: Scenario Creation

Introduction

To improve content specificity of training materials, which has been shown to better capture disorder-specific biases [43,44], living/lived experts were invited to generate CBM materials for paranoia based on their common everyday experiences. We aimed to adapt user-generated scenarios into CBM intervention items.

Method

Participants

Living/lived experts (N = 18) were recruited from the Lived Experience Advisory Panel (LEAP) with the help of co-author Dr Thomas Kabir from the McPin Charity Foundation—an organization based in the UK that focuses on mental health research [47]. McPin collaborates with living/lived experts to invite their feedback in research. Experts were reimbursed for their contribution to this study at £30 per hour. The study has been approved by the London-Stanmore Research Ethics Committee (ref: 21/LO/0896).

Scenario Creation Outline

Intervention Items. We provided our living/lived experts with written information on CBM and guidelines, plus examples for creating exemplars of personal, everyday life scenarios that could provoke paranoid thinking (see supplementary material for a full description).

The STOP research team adapted suitable scenarios (excluding items that were too bizarre, triggering, or did not capture ambiguity) into 240 intervention items in the format commonly used for CBM training items (see Figure 2). Each item consists of three lines of text depicting an emotionally ambiguous scenario that could be either interpreted as paranoid or non-paranoid. The item remains ambiguous until the final word. The final word contains missing letters and is used to resolve the scenario in a non-paranoid manner. One or more

letters (depending on the length of the final word) are removed from the final word (in some items this encompasses the last 2-3 words).
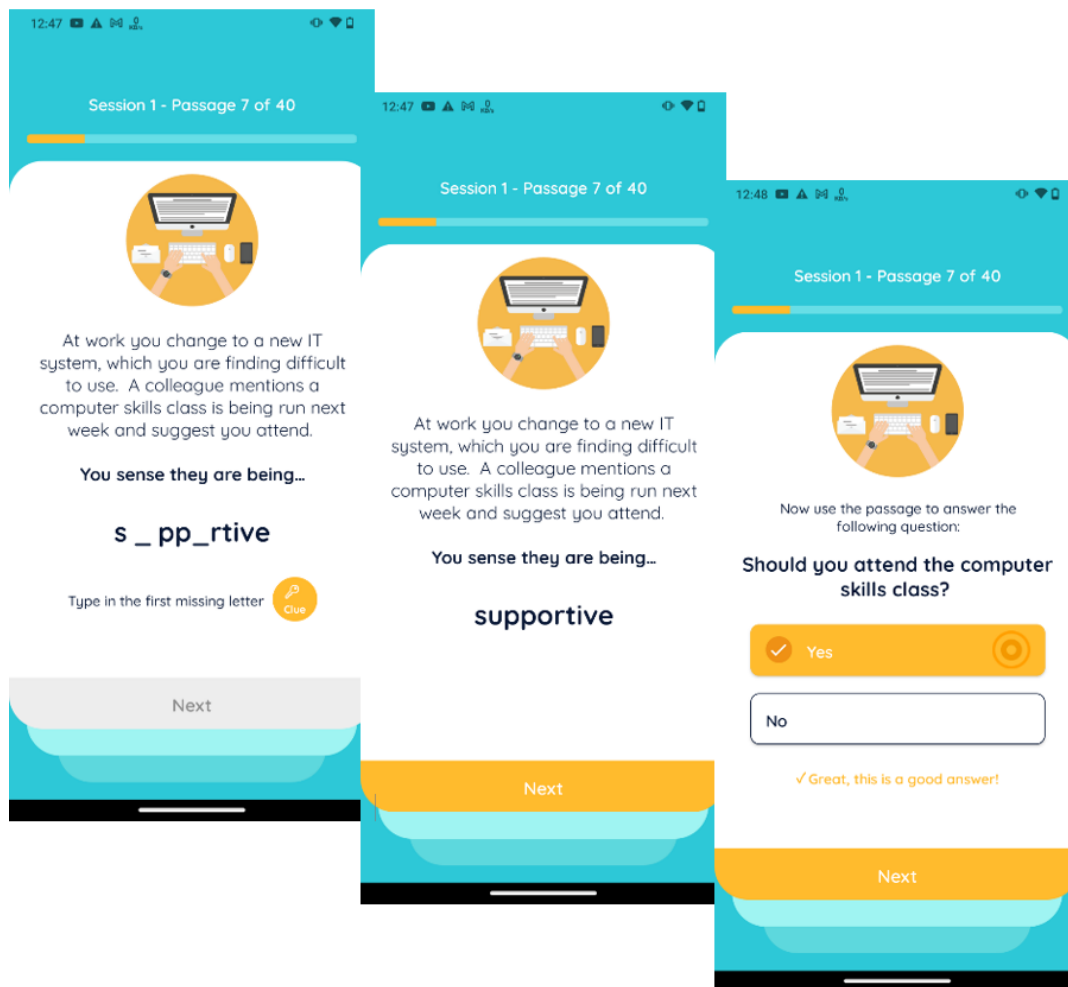

Figure 2. Example of STOP intervention item. Copyright © 2021. Jenny Yiend, King's College London. All rights reserved.

Text-Reading Control. Two hundred and forty control items were created based on non-emotional factual information or mundane activities or sequences of actions (e.g., making a cup of tea). The control items excluded depictions of social situations, emotional words, and feelings. Items were arranged into two topic areas/categories: *general knowledge and facts, and everyday activities.* The format of control items matched that of the intervention items (see Figure 3).
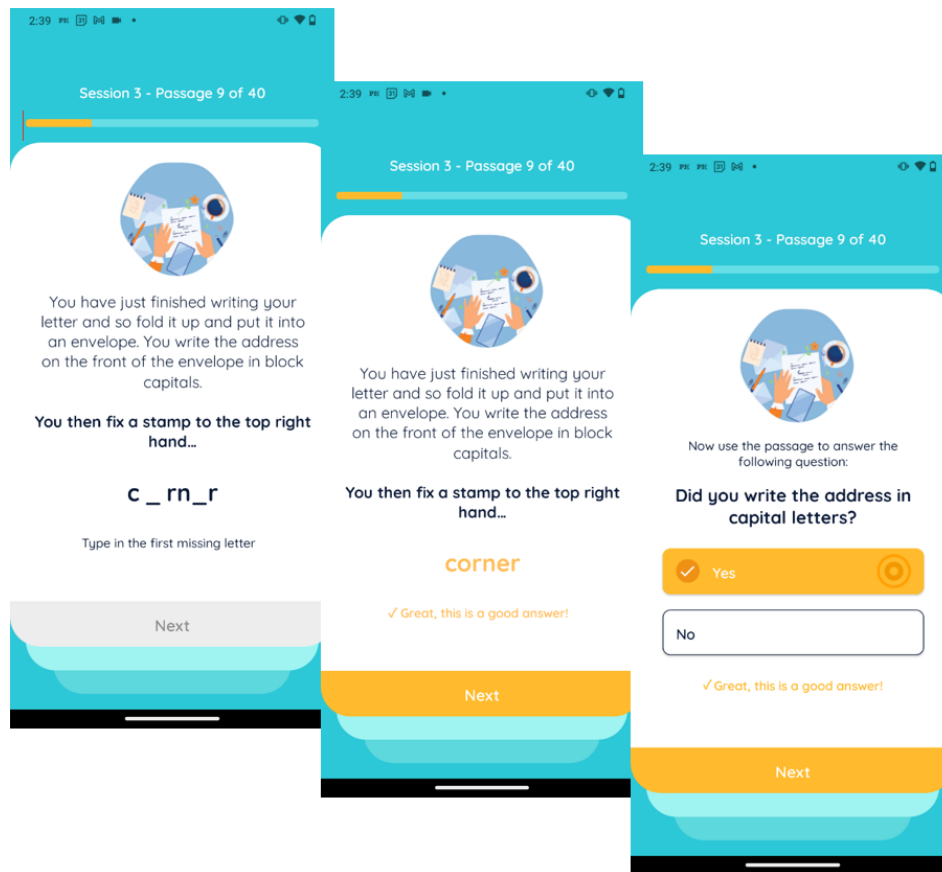
Figure 3. Example of STOP control item. Copyright © 2021. Jenny Yiend, King's College London. All rights reserved.

Stage II: Scenario Evaluation

Introduction

Before employing the items that were created in Stage I as training materials for STOP, these items required further validation to ensure their relevance to and suitability for paranoia. Items were rated for *paranoia severity* and *readability*, and *item length* were recorded. We aimed to reduce systematic discrepancies between intervention and control items and between sessions by matching the *readability of items* and *the item length*. Matching these aspects across intervention versus control item sets and individual weekly sessions within each set may reduce possible confounding effects. For instance, differences in item comprehension or time spent engaging with each item could inadvertently influence the

'dose' of a session. Items were also rated by relevant contributors based on *paranoia severity* with the aim to distribute intervention such that early training sessions included less severe items, with a graded progression toward more potentially threatening/paranoia severity items in later sessions. This was done to allow habituation of any emotional response to occur gradually as the patient progresses through to more challenging therapeutic content, and is usually more acceptable to patients, reducing the risk of dropout. Intervention items also consisted of items with higher paranoia severity ratings compared to the control set.

Method

Participants

We approached a total of 16 raters; half the raters were a group of living/lived experts independent from those who had created the contents in Stage I. Experts were recruited from LEAP. The other half of the raters were clinical psychologists recruited from the Psychological Interventions Clinic for outpatients with Psychosis (PICuP). 15 raters completed all ratings (clinical psychologist: n = 7; living experts: n = 8) due to drop out from being busy after only completing one-third of the ratings. Raters were randomly assigned to rate either intervention (n = 8) or control items (n = 7). Clinician raters and living/lived experts were reimbursed for their contribution to the study at £50 and £30 per hour, respectively.

Procedures

For the purpose of rating, we included the final word of the passage that completes the text and removed the follow-up yes/no question. For the intervention item, the final word depicted the paranoid interpretation of the ambiguous text. Clinician raters rated the intervention scenarios based on the criteria: *paranoia severity and readability.* For example, raters were asked to rate the level of paranoia each scenario is likely to evoke (see Supplementary Materials for additional information on counterbalancing of ratings).

Paranoia severity was rated on a 6-point scale (0=*not paranoid*; 1=*mild paranoia* to 5=*severe paranoia*); readability was rated on a 6-point scale (0 = *difficult to read* 5 = *easy to read*). A mean rating of ≥ 1 for the intervention item and ≤ 1 for the control item was set, a priori, as the acceptable threshold for the severity scale; A mean rating of ≥ 3 was set, a priori, as the acceptable threshold for the readability scale for both experimental conditions. Living/lived experts rated items on the readability criterion only. Paranoia ratings from living experts were not appropriate because to gauge the severity of the potentially paranoid content it was necessary to present items in their negative/ paranoid form. This would be a prolonged, unjustifiable, and potentially harmful negative mood induction for these individuals.

Once all data were collected from raters, we conducted an iterative process of reviewing and refining items. First, means were calculated for paranoia severity and readability. Items that fell below the acceptable value were reviewed or replaced (n = 43 intervention items did not reach the threshold on the severity scale). These items were discussed among the STOP team, re-written, and then re-rated by two of the same clinicians (see Supplementary Materials for interrater reliability data). Finally, three 2-hour Zoom meetings were conducted with 4-6 living/lived experts at each meeting to systematically review, item by item, the final intervention and control content. Feedback was recorded and further minor replacements/revisions were made where essential.

Items were distributed based on paranoia severity, readability, and item length. We evenly distributed intervention items into six, 40-item sessions based on a progression of mean paranoia severity ratings across the six sessions (while checking for any discrepancies between readability ratings between intervention and control item sets and between the six sessions). Item length—operationally defined by the item's total character count—was also matched within and between sessions and item sets (see Supplementary Materials for additional information on cross-referencing of item length).

Results

In the first iteration of re-rating, 24 training items reached acceptable values (Paranoia severity: $M = 3.48$, $SD = .95$), whereupon all items reached the threshold after re-rating (Paranoia severity: $M = 4.71$, $SD = .30$). All control items reached the acceptable value for paranoia severity and readability (see Supplementary Materials for the Analysis Plan). For item distribution based on paranoia severity, as shown in Table 1, a 2 (intervention, control) x 6 (sessions 7-12) ANOVA showed a systematic difference in items' severity between intervention and control item, $F(1, 468) = 6201.01$, $p < .001$, between sessions, $F(5, 468) = 194.76$, $p < .001$, and there was an interaction, $F(5, 468) = 223.07$, $p < .001$. Post hoc examination of the mean severity scores revealed that there was a difference in items' severity across sessions for the intervention but not the control group (see Figure 4). In STOP, the six sessions previously developed as part of the feasibility study [36] were interleaved with the six newly created sessions to create 12 sessions based on a progression of mean paranoia severity ratings.
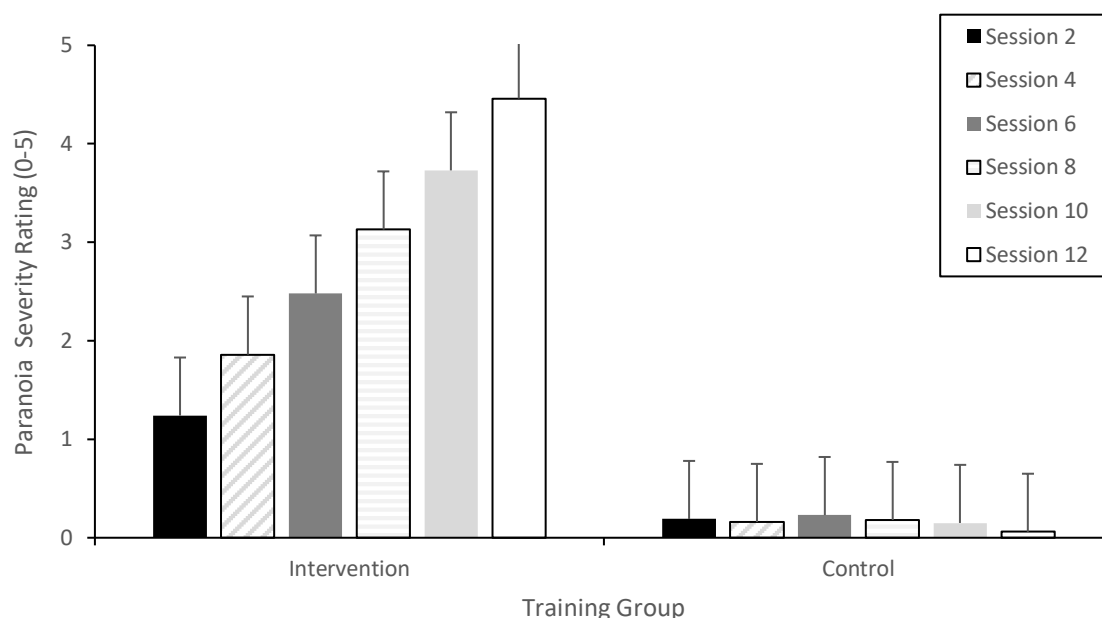

Figure 4. Mean paranoia severity ratings across training groups and sessions.

For item length, cross-checking by three researchers showed a high agreement for both intervention (93%) and control items (94.5%), with 100% agreement between the researchers following resolution. For item distribution based on item length, as shown in Table 1, a 2 (intervention, control) x 6 (sessions 7-12) ANOVA revealed no systematic differences in the item's character count between intervention and control items, $F(1, 468) = 1.43$, $p = .23$, between sessions, $F(5, 468) = .01$, $p = 1.00$, and there was no interaction, $F(5, 468) = .12$, $p = .99$.

Table 1. Mean (SD) character count and item ratings (intervention; control) of paranoia severity and readability across sessions.

Session	Intervention Items - Mean (*SD*)				Control Items - Mean (*SD*)			
	Clinician Rating (n=8)		User Rating (n=8)	Scenario Character Count	Clinician Rating (n=7)		User Rating (n=8)	Scenario Character Count
	Severity	Readability	Readability		Severity	Readability	Readability	
7	1.24	4.17	3.76 (.60)	153.65	.19 (*.23*)	4.23	3.69	156.62
	(*.33*)	(.50)		(*23.06*)		(.51)	(.73)	(*41.24*)
8	1.86	4.21	3.73 (.62)	154.32	.16 (*.24*)	4.21	3.82	155.30
	(*.47*)	(.55)		(*25.18*)		(.65)	(.63)	(*36.61*)
9	2.48	4.29	3.76 (.53)	155.07	.23 (.28)	3.99	3.50	156.28
	(.50)	(.48)		(*30.66*)		(.55)	(.47)	(*40.69*)
10	3.13	4.46	3.82 (.57)	151.62	.18 (*.24*)	4.23	3.53	159.28
	(*.56*)	(.35)		(*28.87*)		(.52)	(.69)	(*34.43*)
11	3.73	4.24	3.76 (.65)	153.73	.15 (*.21*)	4.13	3.55	158.05
	(*.57*)	(.54)		(*28.13*)		(.69)	(.75)	(*32.47*)
12	4.46	4.37	4.10 (.60)	152.95	.06 (*.15*)	4.22	3.56	156.88
	(*.33*)	(.38)		(29.97)		(.64)	(.61)	(*28.84*)
Total	2.82	4.29	3.82 (*.60*)	153.56	0.16 (*.23*)	4.17	3.61	157.07
	(*1.19*)	(.48)		(*27.50*)		(.60)	(.67)	(*35.63*)

Stage III: Item Graphics

Introduction

In the CBM-pa feasibility trial [36], living/lived experts recommended visually enriching content in addition to text passages to increase the engagement and realism of text scenarios [30]. Indeed, researchers have shown that the effectiveness of CBM clinical interventions is positively correlated with the degree of participants' active involvement [48]. We, therefore included graphics to accompany each of the intervention and control items used in STOP.

Method

Materials

Graphics development was outsourced to an industry partner, Avegen [46]. The STOP research team provided Avegen with text-based scenarios that were developed in the previous stages of this study. Avegen graphics designers created the graphics based on extrapolations of the text-based scenarios. The graphics were chosen to depict the ambiguous scenarios and their non-paranoid interpretation (that runs counter to the paranoid reader's initial assumption), as well as the neutral control items. Three types of graphics were included (see Figure 5, for an example of each one: 1) static images (n = 576), 2) dynamic images (n = 192), and 3) scenes (n = 192; each a collection of three static images depicting the sequence of events in the unfolding scenario).
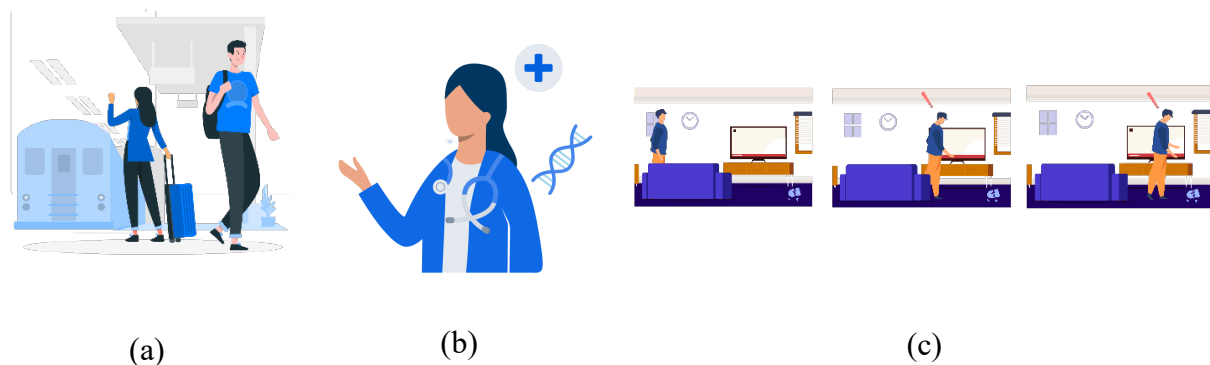

(a)   (b)   (c)

Figure 5. Example of item graphics: a) static image; b) dynamic image; c) 3-image scene

Participants and Procedure

Once graphics were created, we invited 18 unreimbursed volunteers from the community to rate a random selection (totalling one-quarter of all material) of the graphics used in STOP based on specific attributes of user experience. We randomly selected 25% (N = 120) of each type of graphics for the six newly created sessions for STOP (total 480 items), and then randomly assigned half of the users (n = 9) to rate graphics of intervention items and the other half (n = 9) to rate graphics of control items. Participants rated the graphics independently on three rating criteria: *complexity, attractiveness,* and *interest* using three 100-point sliding scales (0 = *the least...* to 100 = *the most...*), one for each rating criterion. The three rating criteria were selected by two researchers from 10 scales of the User Experience Questionnaire that described the appearance of interactive products [49]. The three rating criteria were selected based on coverage of the scales and their relevance to STOP. At the outset of the graphics rating task, we showed users an example of two images on opposite ends of the scales for each rating criterion as anchors. Graphics were presented as a Qualtrics survey with the following instructions, "*Welcome to the rating questionnaire. There are 120 items and it should take around 20-30 minutes. Using the sliders, please rate each of the following images against the parameters below*".

Results

Table 2 shows ratings on training item graphics as a function of item category (Intervention; Control). A series of independent-samples t-tests indicated no significant difference between intervention and control graphics across all three rating scales (*complexity, attractiveness, interesting).*

Table 2. User Rating (Mean and Standard Deviations) on training item graphics (Max score = 100) as a Function of Item Category (Intervention; Control).

	Intervention Items Mean (SD)	Control Items Mean (SD)	t test	P value
Complexity	45.64 (22.54)	46.05 (22.80)	.42	.68
Attractiveness	57.72 (19.46)	56.50 (20.26)	1.43	.15
Interest	56.90 (20.85)	55.57 (21.19)	1.47	.14

Stage IV: STOP Mobile App Usability Testing

Method

STOP App Design

The STOP App development was outsourced to Avegen [46]. STOP is a mobile App that delivers CBM therapy for paranoia on either Android or iOS platforms. In consultation with the STOP research team, Avegen designed and built the App top-down using the finalized training items developed in the previous stages of this work. STOP provides one self-directed weekly therapy session consisting of 40 training items, taking approximately 40 minutes to complete. Users schedule weekly sessions on their STOP phone App; automatic reminders are sent to users via email before the session. Each item includes user-generated text-based scenarios with accompanying graphics. Session content is interspersed with trivia and badges upon completion of each training session to improve user experience. Living experts are invited to test the STOP phone App and provide feedback during two pilot sessions (May 2021; October 2021). Initial aspects of the app design (e.g., STOP acronym, logo design, colour palette, fonts, layout, storyboard, gamification elements, instructions for use etc.) were co-designed with the LEAP group over a period of 6 months through a series of regular group meetings attended by the industry partner and relevant graphic designers.

Once the first MVP (minimal viable product) was achieved the formal phase of usability testing began.

Usability Testing

Participants. A group of living/lived experts (pilot 1: n = 5; pilot 2: n = 4) separate from those who contributed to the previous stages of this work were recruited from LEAP as a part of the usability testing for STOP. Again, living experts were reimbursed for their contribution to this study at £30 per hour.

Procedures. Two piloting sessions of the STOP mobile application were scheduled with living experts to incorporate feedback to refine and improve the product. The first pilot study lasting approximately 45 minutes included a *test version* of STOP in that the content and function of the App were limited; the second pilot study included the testing of two intervention sessions across two weeks (from 11 October 2021 to 22 October 2021). In both pilot studies, living experts provided quantitative ratings on the following features of the mobile app: ease of use, user interface, interactive features, design and graphics, security and privacy, errors/bugs, help provision (See Supplementary Materials for a description of each feature). These criteria were adapted from the User Experience Questionnaire [49]. Living/lived experts provided a rating of each feature using a 5-point scale (*1= inadequate, 2= adequate, 3=good, 4=very good, 5=excellent*). A mean rating of ≥ 2 was set, a priori, as the acceptable threshold for each scale.

In addition to the ratings described above, in Pilot 2, we wanted to understand the kinds of problems/issues users were experiencing and their general experience with the STOP mobile application. As such, we invited users to provide a descriptive account of their experience (e.g., *"In one or two sentences, describe any problems/issues that you might have encountered when using the App, if any." "In one or two sentences, describe your overall experience with the App and what you would change, if any"*).

Results

Table 3 shows users' ratings of the STOP mobile application in Pilot 1 and Pilot 2 (see Supplementary Material for users' descriptive accounts). As shown in Table 3, in both pilots, living/lived experts provided a mean rating above our acceptable threshold for all the evaluated features of the STOP mobile application.

Table 3. Mean (SD) user ratings of the STOP mobile phone app (Max score = 5) from usability testing.

STOP Mobile App Feature	Pilot 1	Pilot 2
Ease of use	4.2 (.45)	4.25 (.50)
User interface	4.2 (.84)	4.5 (.58)
Interactive features	4.2 (.84)	4.25 (.96)
Design and graphics	4.4 (.55)	5 (.00)
Help provision	*N/A*	4 (.82)
Security and privacy	4.4 (.55)	5 (.00)
Errors/bugs	4.4 (.89)	*Descriptive account of any errors/bugs (see Table 3b). Any errors identified have been resolved.*
Overall experience	4 (.71)	*Descriptive account of overall experience with the App (see Supplementary Material)*

Discussion

The present study focused on the development of new material to be used in STOP—a novel mobile phone application designed to reduce symptoms of paranoia. This self-administered digital therapeutic aims to reduce symptoms by presenting everyday ambiguous situations that can trigger paranoid thoughts and then normalizing users' interpretations of these situations. However strong the conceptual basis of a new therapeutic, its quality, acceptability, and efficacy will be dependent upon its detailed content and input and recommendations from various relevant contributors [39,41]. This is especially true of interventions that are based upon CBM methods, which rely solely on content for their effect

[25,26], and interventions that address psychosis [38,40,41]. The work presented in the present paper represents a twelve-month activity with clinicians, living/lived experts, a digital solutions design team, and researchers, to develop and evaluate therapeutic content and mobile application of STOP. Specifically, the co-design approach represents a thorough attempt to achieve our two objectives: 1) to take a user-centered approach to create and evaluate paranoia-relevant CBM item content and 2) to engage with living/lived experts and the digital solutions design team to create and pilot test the STOP mobile app prototype.

For all training materials, we reached a priori-defined acceptable threshold for all rating criteria: *paranoia severity and readability of the scenarios,* and there were no systematic differences in item length between intervention and control content, nor within the six newly created sessions of STOP. These data were used to inform the progression of the therapeutic intervention by arranging session content in order of increasing paranoia severity. To reflect clinician-administered cognitive therapies, a 'drill-down' approach from surface-level automatic thoughts to more profound core beliefs was adopted across sessions by using selected specific verbs to reflect each level of thought process. For item graphics, we also found no systematic differences in users' ratings of *complexity, attractiveness, and interest*, between intervention and control groups. Furthermore, evaluations from two pilot tests of STOP with living/lived experts showed that user ratings were above our a priori acceptable thresholds for all evaluated features of the mobile application, suggesting that users found the STOP App easy to use and navigate, suitably interactive, helpful, and secure.

The existing literature demonstrates the importance of co-designing mobile phone Apps for psychosis with multiple collaborators [38,39,40]. Our present work illustrates one approach to implementing a detailed user-centered development process that was applied throughout the entire design and development process of a mobile application. This may serve as a useful model for others, as the field of digital mental health continues to grow

exponentially. Our co-design is likely to have improved the relevance, authenticity, face validity, and acceptability of both the therapy interface and its content, compared to a researcher-led approach, although we cannot provide direct evidence of this. In each phase of STOP's creation, we involved relevant contributors to provide feedback, open discussions, and formal usability testing of STOP's content and mobile App. Contrary to STOP's predecessor CBM-pa [36] where only the researchers designed training materials, in our present work, we refined both the therapeutic content (training material) and the mobile App implementation, following contributors' recommendations. The literature on co-design suggests that the careful and inclusive development process we have followed is likely to enhance user engagement and uptake of STOP [39]. There is also evidence that co-design improves treatment adherence and motivation [51]. Further additional features that we have included, such as graphical enhancement, use of therapeutic content based on actual patient experience, and close attention to the reduction in potential confounding variables (e.g., time spent in therapy could inadvertently influence the 'dose' of a session), may improve the intervention when tested against a control group in a clinical trial.

Several improvements could be made to the present study, however. First, despite basing content development on user-generated examples, the personal relevance to any one individual is limited. Future work should consider ways to tailor content to the individual in real-time or prior to the start of therapy. The development of personalised predictive algorithms and agile methods of therapeutic content selection will be one way to do this. Second, it will be important to test the STOP app for acceptability and feasibility of usage in a live clinical service setting, as called for by national organisations such as NICE [50]. By the same token, our small sample of raters was recruited from single clinical service units within the United Kingdom, thereby limiting the representativeness of the feedback and ratings received.

Third, there are several limitations relevant specifically to a clinical trial context use of STOP, as opposed to real-world deployment. For example, we only matched items by length between experimental and control groups, as measured by items' character counts; a more thorough matching process would yield greater confidence that when testing STOP against its active control any differences could not be the result of unintended material effects. Using single factors such as these to control for arbitrary effects of the intervention is limited, and in the future, other factors could be added to better control for confounds (e.g., measuring actual reading speed, user's comprehension of items, gender-specific content, intercultural relevance and so on).

A further trial-related limitation is that graphics were rated on only three rating criteria pertaining to visual appearance, which were derived from subscales of a standardised instrument. The limited selection of scales was a pragmatic decision and future work could match graphic content on a wider range of criteria, for example including aspects of appearance, such as aesthetics, excitement, likeability, and others, all of which are included in the original instrument that was used to motivate our selection of scales. In addition to graphic enrichments, other elements, including badges, progress trackers and trivia are integral to the STOP mobile app and are derived from earlier focus group discussions, but these have not been evaluated. Ideally, all enrichments should be tested systematically to determine their effectiveness in engaging and motivating service users.

Finally, although we rely on feasibility data and previous ratings and feedback [26] to validate the first six sessions of STOP, nevertheless, an improvement in future work would be to evaluate all 12 sessions simultaneously on the same metrics.

Conclusion

In conclusion, CBM-pa is a relatively recent novel psychological intervention that has now been extended into the digital therapeutic called STOP. Material development and App

design for any new CBM content should follow an iterative and rigorous process involving

multiple contributors, including living/lived experts, researchers, clinicians, and the design

team. This user-centered approach to intervention development maximizes the relevance of

therapeutic content to the target user group. In so doing, researchers will most likely also

optimize user acceptability, effectiveness, and engagement to create the best possible mobile

health interventions for people with severe psychiatric disorders.

Ethics. The STOP trial program of work received ethical approval from the London–

Stanmore Research Ethics Committee (reference: 21/LO/0896) and all those participating in

the work described gave consent for publication.

Conflict of Interest. The authors have no conflicts of interest to declare.

References

[1] Freeman, D., McManus, S., Brugha, T., Meltzer, H., Jenkins, R., & Bebbington, P. (2011). Concomitants of paranoia in the general population. *Psychological medicine*, *41*(5), 923-936. PMID: 20735884

[2] Schizophrenia Commission. (2012). The abandoned illness: a report from the Schizophrenia Commission (Rethink Mental Illness, London). PMID: 24354212

[3] Freeman, D., Garety, P. A., Kuipers, E., Fowler, D., & Bebbington, P. E. (2002). A cognitive model of persecutory delusions. *British Journal of Clinical Psychology*, *41*(4), 331-347. PMID: 12437789

[4] Wessely S., Buchanan, A., Reed, A., Cutting, J., Everitt, B., Garety, P., & Taylor, P. J. (1993). Acting on delusions. I: Prevalence. *The British Journal of Psychiatry, 163*(1), 69–76. PMID: 8353703

[5] Castle, D. J., Phelan, M., Wessely, S., & Murray, R. M. (1994). Which patients with non-affective functional psychosis are not admitted at first psychiatric contact? *The British Journal of Psychiatry, 165*(1), 101–106. PMID: 7953011

[6] Johnson, J., Horwath, E., & Weissman, M.M. (1991). The validity of major depression with psychotic features based on a community study. *Archives of General Psychiatry, 48*(12), 1075–1081. PMID: 1845225

[7] Goodwin, F. K., & Jamison, K. R. (2007). *Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression*. New York: Oxford University Press. PMID: 22706590

[8] Hamner M. B., Frueh B. C., Ulmer H. G., & Arana G. W. (1999). Psychotic features and illness severity in combat veterans with chronic posttraumatic stress disorder. *Biological Psychiatry, 45*(7), 846–852. PMID: 10202572

[9] van OS, J., Verdoux, H., Maurice-Tison, S., Gay, B., Liraud F., Salamon, R., & Bourgeois, M. (1999). Self-reported psychosis-like symptoms and the continuum of

psychosis. *Social Psychiatry and Psychiatric Epidemiology, 34*(9), 459–463. PMID: 10541665

[10] Appelbaum, P. S., Robbins, P. C., & Roth, L. H., (1999). Dimensional approach to delusions: comparison across types and diagnoses. *The American Journal of Psychiatry*, *156*(12), 1938–1943. PMID: 10588408

[11] Mehl, S., Werner, D., & Lincoln, T. M. (2015). Does Cognitive Behavior Therapy for psychosis (CBTp) show a sustainable effect on delusions? A meta-analysis. *Frontiers in Psychology, 6*, article: 1450. PMIDL 26500570

[12] van der Gaag, M., Valmaggia, L. R., & Smit, F. (2014). The effects of individually tailored formulation-based cognitive behaviour therapy in auditory hallucinations and delusions: a meta-analysis. *Schizophrenia Research, 156*(1), 30–37. PMID: 24737619

[13] Lincoln, T., & Peters, E. (2019. A systematic review and discussion of symptom specific cognitive behavioural approaches to delusions and hallucinations. *Schizophrenia Research, 203,* 66–79. PMID: 29352708

[14] Freeman, D., Emsley, R., Diamond, R., Collett, N., Bold, E., Chadwick, E., Isham, L., Bird, J. C., Edwards, D., Kingdon, D., Fitzpatrick, R., Kabir, T., Waite, F., Oxford Cognitive Approaches to Psychosis Trial Study Group. Comparison of a theoretically driven cognitive therapy (the Feeling Safe Programme) with befriending for the treatment of persistent persecutory delusions: a parallel, single-blind, randomised controlled trial. *Lancet Psychiatry, 8*(8), 696–707. PMID: 34246324

[15] Elkis, H. (2007). Treatment-resistant schizophrenia. *Psychiatric Clinics of North America*, *30*(3), 511–533. PMID: 17720034

[16] Turner D. T., van der Gaag, M. Karyotaki, E., & Cuijpers, P. (2014). Psychological interventions for psychosis: a meta-analysis of comparative outcomes studies. *The American Journal of Psychiatry, 171*(5), 523–538. PMID: 24525715

[17] Moritz, S., & Woodward, T. S. (2007). Metacognitive training in schizophrenia: from basic research to knowledge translation and intervention. *Current Opinion in Psychiatry, 20*(6), 619–625. PMID 17921766

[18] Waller, H., Freeman, D., Jolley, S., Dunn, G., & Garety, P. (2011). Targeting reasoning biases in delusions: a pilot study of the Maudsley Review Training Programme for individuals with persistent, high conviction delusions. *The Journal of Behavior Therapy and Experimental Psychiatry, 42*(3), 414–421. PMID: 21481815

[19] Chen, J., Short, M., & Kemps, E. (2020). Interpretation bias in social anxiety: A systematic review and meta-analysis. *Journal of Affective Disorders*. PMID: 32777650

[20] Everaert, J., Podina, I. R., & Koster, E. H. (2017). A comprehensive meta-analysis of interpretation biases in depression. *Clinical Psychology Review*, *58*, 33-48. PMID: 28974339

[21] Amin, N., Foa, E. B., & Coles, M. E. (1998). Negative interpretation bias in social phobia. *Behaviour research and therapy*, *36*(10), 945-957. PMID: 9714945

[22] Freeman, D., Garety, P. A., Kuipers, E., Fowler, D., & Bebbington, P. E. (2002). A cognitive model of persecutory delusions. *British Journal of Clinical Psychology, 41*(pt4), 331–347. PMID: 12437789

[23] Yiend, J., Allen, P., Lopez, N. D., Falkenberg, I., Tseng, H. H., & McGuire, P. (2019). Negative interpretation biases precede the onset of psychosis. *Behaviour therapy*, *50*(4), 718-731. PMID: 31208682

[24] Trotta, A., Kang, J., Stahl, D., & Yiend, J. (2021). Interpretation bias in paranoia: a systematic review and meta-analysis. *Clinical Psychological Science*, *9*(1), 3-23. Doi: 10.1177/2167702620951552

[25] Savulich, G., Freeman, D., Shergill, S., & Yiend, J. (2015). Interpretation biases in paranoia. *Behaviour therapy*, *46*(1), 110-124. PMID: 25526839

[26] Savulich, G., Shergill, S. S., & Yiend, J. (2017). Interpretation biases in clinical paranoia. *Clinical Psychological Science*, *5*(6), 985-1000. Doi: 10.1177/2167702617718180

[27] Cristea, I. A., Kok, R. N., & Cuijpers, P. (2015). Efficacy of cognitive bias modification interventions in anxiety and depression: meta-analysis. *The British Journal of Psychiatry*, *206*(1), 7-16. PMID: 25561486

[28] Cristea, I. A., Kok, R. N., & Cuijpers, P. (2016). The effectiveness of cognitive bias modification interventions for substance addictions: a meta-analysis. *PloS one*, *11*(9), e0162226. PMID: 27611692

[29] Jones, E. B., & Sharpe, L. (2017). Cognitive bias modification: A review of meta-analyses. *Journal of affective disorders*, *223*, 175-183. PMID: 28759865

[30] Leung, C. J., Fosuaah, A., Frerichs, J., Heslin, M., Kabir, T., Lee, T. M. C., ... & Yiend, J. (2019). A qualitative study of the acceptability of cognitive bias modification for paranoia (CBM-pa) in patients with psychosis. *BMC psychiatry*, *19*(1), 1-13. PMID: 31337373

[31] Malla, A. K., Norman, R. M. G., Manchanda, R., & Townsend, L. (2002). Symptoms, cognition, treatment adherence and functional outcome in first-episode psychosis. *Psychological Medicine*, *32*(6), 1109-1119. PMID: 12214790

[32] Yang, R., Cui, L., Li, F., Xiao, J., Zhang, Q., & Oei, T. P. S. (2017). Effects of cognitive bias modification training via smartphones. *Frontiers in Psychology, 8*, 1370. PMID: 28855880

[33] Beard, C., Ramadurai, R., McHugh, R. K., Pollak, J. P., & Björgvinsson, T. (2021). HabitWorks: Development of a CBM-I smartphone app to augment and extend acute treatment. *Behavior Therapy, 52*(2), 365–378. PMID: 33622506

[34] Steel, C., Wykes, T., Ruddle, A., Smith, G., Shah, D. M., & Holmes, E. A. (2010). Can we harness computerised cognitive bias modification to treat anxiety in schizophrenia? A first step highlighting the role of mental imagery. *Psychiatry Research*, *178*(3), 451-455. Doi: 10.1016/j.psychres.2010.04.042

[35] Turner, R., Hoppitt, L., Hodgekins, J., Wilkinson, J., Mackintosh, B., & Fowler, D. (2011). Cognitive bias modification in the treatment of social anxiety in early psychosis: a single case series. *Behavioural and Cognitive Psychotherapy*, *39*(3), 341-347. PMID: 21320359

[36] Yiend, J., Lam, C. L. M., Schmidt, N., Crane, B., Heslin, M., Kabir, T., McGuire, P., Meek, C., Mouchlianitis, E., Peters, E., Stahl, D., Trotta, A., & Shergill, S. (2022). Cognitive bias modification for paranoia (CBM-pa): a randomised controlled feasibility study in patients with distressing paranoid beliefs. *Psychological Medicine, 14*, 1–13. PMID: 35699135

[37] Nilsen, W., Kumar, S., Shar, A., Varoquiers, C., Wiley, T., Riley, W. T., & Atienza, A. A. (2012). Advancing the science of mHealth. *Journal of Health Communication, 17*(sup1), 5–10. PMID: 22548593

[38] Ben-Zeev, D., Brenner, C. J., Begale, M., Duffecy, J., Mohr, D. C., & Mueser, K. T. (2014). Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. *Schizophrenia Bulletin, 40*(6), 1244–1253. PMID: 24609454

[39] Schouten, S. E., Kip, H., Dekkers, T., Deenik, J., Beerlage-de Jong, N., Ludden, G. D. S., Kelders, S. M. (2022). Best-practices for co-design processes involving people with severe mental illness for eMental health interventions: a qualitative multi-method approach. *Design for Health, 6*(3), 316–344. Doi: 10.1080/24735132.2022.2145814

[40] Kidd, S.A., Feldcamp, L., Adler, A., Kaleis, L., Wang, W., Vichnevetski, K., et al. (2019). Feasibility and outcomes of a multi-function mobile health approach for the

schizophrenia spectrum: app4Independence (A4i). *PLoS One, 14*(7): e0219491. PMID: 31306439.

[41] Berry, N., Machin, M., Ainsworth, J., Berry, K., Edge, D., Haddock, G., Lewis, S., Morris, R., & Bucci, S. (2020). Developing a theory-informed smartphone app for early psychosis: Learning points from a multidisciplinary collaboration. *Frontier in Psychiatry, 11*: 602861. PMID 3336212.

[42] Maguire, M. (2001). Methods to support human-centered design. *International Journal of Human-Computer Studies, 55*(4), 587–634. Doi: 10.1006/ijhc.2001.0503

[43] Mathews, A., & MacLeod, C. (1994). Cognitive approaches to emotion and emotional disorders. *Annual review of psychology*, *45*(1), 25-50. PMID: 8135504

[44] Yiend, J., Barnicot, K., Williams, M., & Fox, E. (2018). The influence of positive and negative affect on emotional attention. *Journal of Behavior Therapy and Experimental Psychiatry, 61,* 80–86. PMID: 29990682

[45] Freeman, D., Garety, P. A., Bebbington, P. E., Smith, B., Rollinson, R., Fowler, D., Kuipers, E., Ray, K., & Dunn, G. (2005). Psychological investigation of the structure of paranoia in a non-clinical population. *British Journal of Psychiatry, 186*(5)*,* 427–435. PMID: 15863749

[46] Avegen (2022). https://avegenhealth.com/

[47] McPin Foundation (2018). https://mcpin.org/

[48] Hoppitt, L., Mathews, A., Yiend, J., & Mackintosh, B. (2010). Cognitive bias modification: The critical role of active training in modifying emotional responses. *Behavior Therapy, 41*(1), 73–81. PMID: 20171329

[49] Laugwitz, B., Held, T., Schrepp, M. (2008, November 20-21). *Construction and evaluation of a User Experience Questionnaire* [Conference paper]. HCI and Usability for Education and Work, 4th Symposium of the Workgroup Human-Computer

Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria. Doi: 10.1007/978-3-540-89350-9_6

[50] National Institute for Health and Care Excellence (2019). *Evidence standards framework for digital health technologies.* https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf

[51] Killikelly, C., He, Z., Reeder, C., & Wykes, T. (2017). Improving adherence to web-based and mobile technologies for people with psychosis: systematic review of new potential predictors of adherence. *JMIR Mhealth Uhealth, 5*(7), e94. PMID: 28729235

Abbreviations

Abbreviations	Meaning	Page
CBM-pa	Cognitive Bias Modification-Paranoia	2
STOP	Successful Treatment of Paranoia	2
LEAP	Lived Experience Advisory Panel	10
ANOVA	Analysis of Variance	16
MRC	Medical Research Council	27
DPFS	Developmental Pathway Funding Scheme	27