informatics

MDPI

Article

A Structured Mixed-Methods Workflow for User-Centred Digital Mental Health Platform Design

Harold Ngabo-Woods [1], Isabel Seguí Verdú [2,*], Larisa Dunai [2] and Eva María Burguera Sierra [2]

[1] Doctoral School, Universitat Politècnica de València, 46022 Valencia, Spain; hawooka@doctor.upv.es
[2] Department of Graphic Engineering, Universitat Politècnica de València, 46022 Valencia, Spain; ladu@upv.es (L.D.); evbursie@degi.upv.es (E.M.B.S.)
* Correspondence: issever@doctor.upv.es

Abstract

Digital mental health interventions (DMHIs) frequently experience low user engagement and high attrition, highlighting the importance of incorporating user-centred design approaches during early-stage development. This study presents a structured mixed-methods workflow that integrates participatory co-design activities with a complementary quantitative prototype evaluation to support the formative development of digital mental health platforms. Using a qualitatively driven sequential design with an embedded quantitative strand, 45 university students in design, engineering, architecture and ergonomics programmes took part in three participatory co-design workshops in nine groups of five and, at the close of the final session, each group rated the low-fidelity AMGR-I prototype it had itself built against eleven design criteria. Ratings differed systematically across criteria (Friedman $\chi^2(10) = 225.73$, $p < 0.001$, Kendall's $W = 0.50$), and the difference persisted when the clustering of ratings within the nine artefacts was modelled. Gamification was rated significantly lower than every other criterion after Holm's correction and content quality highest, while data privacy and security was among the lowest-rated criteria despite being one of the most frequently and forcefully discussed concerns in the workshops. Because the instrument measured evaluation of the artefact rather than the importance attached to each criterion, that pattern is a convergence between the two strands rather than a paradox: the requirement these participants discussed most was the one their own artefacts served worst. That convergence is the study's central integrated finding. The participants are characterised throughout as design-literate formative evaluators rather than as prospective users, and the resulting design propositions are presented as hypotheses for subsequent testing. The contribution of the paper is the workflow itself: a lightweight, low-cost and classroom-deployable procedure for combining co-design with structured prototype evaluation at the pre-clinical formative stage.

Keywords: digital mental health; co-design; human–computer interaction; user-centred design; mixed-methods research; user experience

Check for updates

Received: 29 July 2026
Revised: 5 October 2026
Accepted: 20 October 2026
Published: 27 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Digital mental health interventions (DMHIs) have gained significant traction as an accessible means of delivering support, yet their real-world effectiveness is often hampered by poor user engagement and high attrition rates [1–3]. Many digital solutions focus primarily on functional requirements, with less emphasis on creating the engaging and emotionally resonant experiences necessary for sustained use [4,5]. This challenge is

particularly acute among university students, a population that faces significant mental health challenges while also being prolific users of digital technology, making them a relevant population for exploring early-stage digital mental health platform design and evaluation [6–10].

To address this engagement gap, scholars have increasingly advocated for user-centred, co-design approaches, wherein intended users and other stakeholders collaboratively inform technology development from its early stages [11–13]. By involving users directly, interventions can be made more relevant and responsive to their lived experiences, which is theorised to improve adoption and effectiveness [14–17]. This approach ensures that the final product resonates with the diverse backgrounds, cultural contexts, and clinical histories that affect how individuals perceive technology-based interventions [5,18].

However, a persistent challenge within co-design research is its frequent reliance on small-scale, qualitative methods such as workshops and focus groups. While these methods provide rich contextual understanding and detailed user insights, a complementary quantitative strand may support the structured evaluation of a design artefact and a broader interpretation of emerging design themes [19–21]. This creates an applied development challenge: how to retain the richness of qualitative co-creation while integrating quantitative feedback capable of supporting evidence-informed design decisions during early-stage platform development [22].

This paper addresses that challenge by presenting and evaluating a structured mixed-methods workflow for the formative development of DMHIs. Using the creation of the AMGR-I prototype as a case study, we illustrate how participatory co-design workshops can be combined with a structured evaluation of the resulting artefact so that in-depth participant perspectives and broader patterns of evaluation can be interpreted together. Rather than proposing a new theory of co-design, the paper contributes an applied workflow for integrating qualitative and quantitative evaluation within early-stage DMHI development, and reports honestly on what such a workflow can and cannot establish at this scale.

An earlier, single-authored account of the same study was posted as a preprint in November 2023 [23]. The present paper extends that account with the full analytic procedure, the integration mechanism, inferential comparison of the criteria, a formal assessment of factorability, and the workflow formulation that constitutes the contribution.

2. Background and Related Work

2.1. Co-Design in Digital Health

Co-design, which is closely related to participatory design approaches, has proven effective for technologies requiring intensive user involvement, such as mental health interventions [24,25]. In these contexts, involving intended users and relevant stakeholders (e.g., patients, clinicians, caregivers) ensures that the product resonates with their lived experiences [14]. Co-design typically involves iterative cycles of insight gathering, prototype development, and feedback loops, placing user voices at every stage of the design process [20,26–28].

Despite documented benefits, implementing co-design in digital mental health raises unique challenges. First, mental health stakeholders present heterogeneous needs and conditions, which can complicate universal design decisions [29]. Second, bridging clinical expertise with lay user experiences requires careful facilitation [30]. Third, many digital health initiatives adopt singular qualitative approaches, making it more difficult to compare emerging requirements across broader participant groups [31]. Combining qualitative and quantitative methods may support more structured interpretation of design outcomes during early-stage platform development.

2.2. Existing Frameworks and the Gap Addressed Here

Several established frameworks already govern the development and evaluation of digital health interventions, and the contribution claimed here must be stated relative to them rather than in isolation.

The Person-Based Approach [32] is the closest antecedent: it likewise privileges iterative qualitative engagement to surface the beliefs and contexts that shape intervention use, and it likewise operates at the formative stage. The Medical Research Council and NIHR guidance on developing and evaluating complex interventions [33] provides the overarching methodological architecture within which formative work sits, spanning development, feasibility, evaluation and implementation. The NASSS framework [34] addresses the determinants of adoption, non-adoption and abandonment once a technology reaches a care setting, and the Accelerated Creation-to-Sustainment model [35] addresses the compression of development and evaluation cycles for behavioural intervention technologies. Independently, the health preference elicitation literature offers mature instruments, including discrete choice experiments and best–worst scaling, for establishing priorities among competing attributes under explicit trade-off [36,37].

The workflow presented here does not compete with any of these, and it would be misleading to present it as doing so. The Person-Based Approach and the MRC/NIHR guidance operate at a programme scale that presumes access to clinical populations, ethical infrastructure and sustained funding. NASSS presumes a technology mature enough to be adopted or abandoned. Discrete choice experiments and best–worst scaling require instrument design expertise and larger samples than a formative design study typically commands. What is missing between them is a procedure for the earliest pre-clinical stage: the point at which a design team has a concept and no product, needs to know which attributes to build first, and has access to a room of design-literate participants for a handful of sessions.

The contribution claimed here is therefore deliberately modest in scope and specific in kind. It is a lightweight, low-cost and classroom-deployable operationalisation for early formative work, specifying how co-design outputs and a structured evaluation instrument can be generated, analysed separately and integrated through an explicit joint display, together with an honest account of the inferential limits of doing so at this scale. It is a complement to the frameworks above, and a candidate first step feeding into them, not an alternative.

2.3. Integrating Qualitative and Quantitative Methods in Co-Design

An emerging strand of research underscores the value of combining rich qualitative data with complementary quantitative evaluation measures [22,38–40]. Qualitative engagement, through interviews, workshops, or diaries, captures lived experiences, emotional needs, and use-case complexities [21]. Structured quantitative instruments can summarise how a group evaluates a set of defined attributes and can make disagreement between participants visible in a form that group discussion tends to obscure.

When integrated iteratively, these approaches can strengthen co-design:

1. *Making the distribution of views visible:* Aggregated ratings show whether a recommendation voiced strongly by a few participants is shared across the group, which unstructured discussion rarely reveals.
2. *Uncovering nuances:* Qualitative insights can explain anomalies in quantitative findings, revealing the reasoning behind an otherwise puzzling pattern [31].
3. *Driving iteration:* Quantitative results from formative testing can prompt further qualitative investigation, producing a feedback loop that refines prototypes and design decisions [41].

2.4. The Evolving Landscape of User Expectations

Contemporary users of digital products increasingly expect not only functional tools but also experiences offering emotional engagement and personal resonance [41,42]. In mental health contexts, this translates to interest in features that foster community, personalisation, and autonomy, rather than static content delivery alone. For young adults in particular, peer support and a sense of connection can be powerful motivators [43–46]. Recent reviews also highlight that personalisation is a key driver of engagement, though the mechanisms for achieving it (e.g., user choice versus machine learning) are still being explored [1,47,48]. User-centred and participatory design approaches are well suited to addressing these evolving expectations by eliciting both what participants do and why they do it [5,49].

3. Methods

3.1. Design and Overall Workflow

This study employed a qualitatively driven sequential mixed-methods design with an embedded quantitative strand, following the typology of Creswell and Plano Clark [50] and the integration guidance of Fetters and colleagues [51]. The qualitative strand was dominant and determined the study's analytic direction; the quantitative strand was embedded within it, deriving its content from the first workshop and its data from the close of the final workshop.

We note this classification explicitly because the earlier account of this study described the design as convergent parallel. That description was inaccurate. The two strands were neither independent nor concurrent: the eleven evaluation criteria were derived from themes generated in Workshop 1, and the rating instrument was administered at the end of Workshop 3, after three sessions of group discussion on the same topics. The consequences of this sequence for interpretation are treated as a primary limitation in Section 5.4 rather than a secondary caveat.

Figure 1 presents the workflow as implemented, showing the phases, their inputs and outputs, the decision points, and the single moment at which the two strands are integrated.

The study received ethical approval from the Ethics Committee of the Blithe Research Initiative (approval number BRI.581506.18215/1/1, issued 18 May 2023). Data were collected in June 2023. The Blithe Research Initiative is affiliated with the study's funder, BGDV-Hanbo Research, as a privately funded research body; this relationship is disclosed in the Conflicts of Interest statement. The funder had no role in the design of the study, in the collection, analysis or interpretation of data, or in the decision to publish.

3.2. Participant Recruitment and Characteristics

A total of 45 participants were recruited from a university in Spain using a purposive sampling strategy. Recruitment materials, including flyers and emails distributed through departmental channels, targeted students enrolled in degree programmes related to design, engineering, architecture, and ergonomics. The sample comprised the whole of the cohort available to the study, and no a priori power analysis was conducted; none of the analyses reported below rests on an assumption of adequate power. The earlier account of this study reported an a priori power calculation for a paired-samples comparison of metrics before and after a design change. No such comparison was undertaken in this study, that calculation did not govern its design, and it is not relied upon here. This strategy was selected to recruit participants who, while not representing a clinical population with lived experience of mental ill-health, possessed a high degree of technological literacy and a developed vocabulary for articulating feedback on design, usability, and user experience. Participants are therefore characterised throughout this paper as design-literate formative evaluators rather than as users or prospective users of a digital mental health platform,

and this characterisation constrains the interpretation of every finding reported below, not only the limitations.

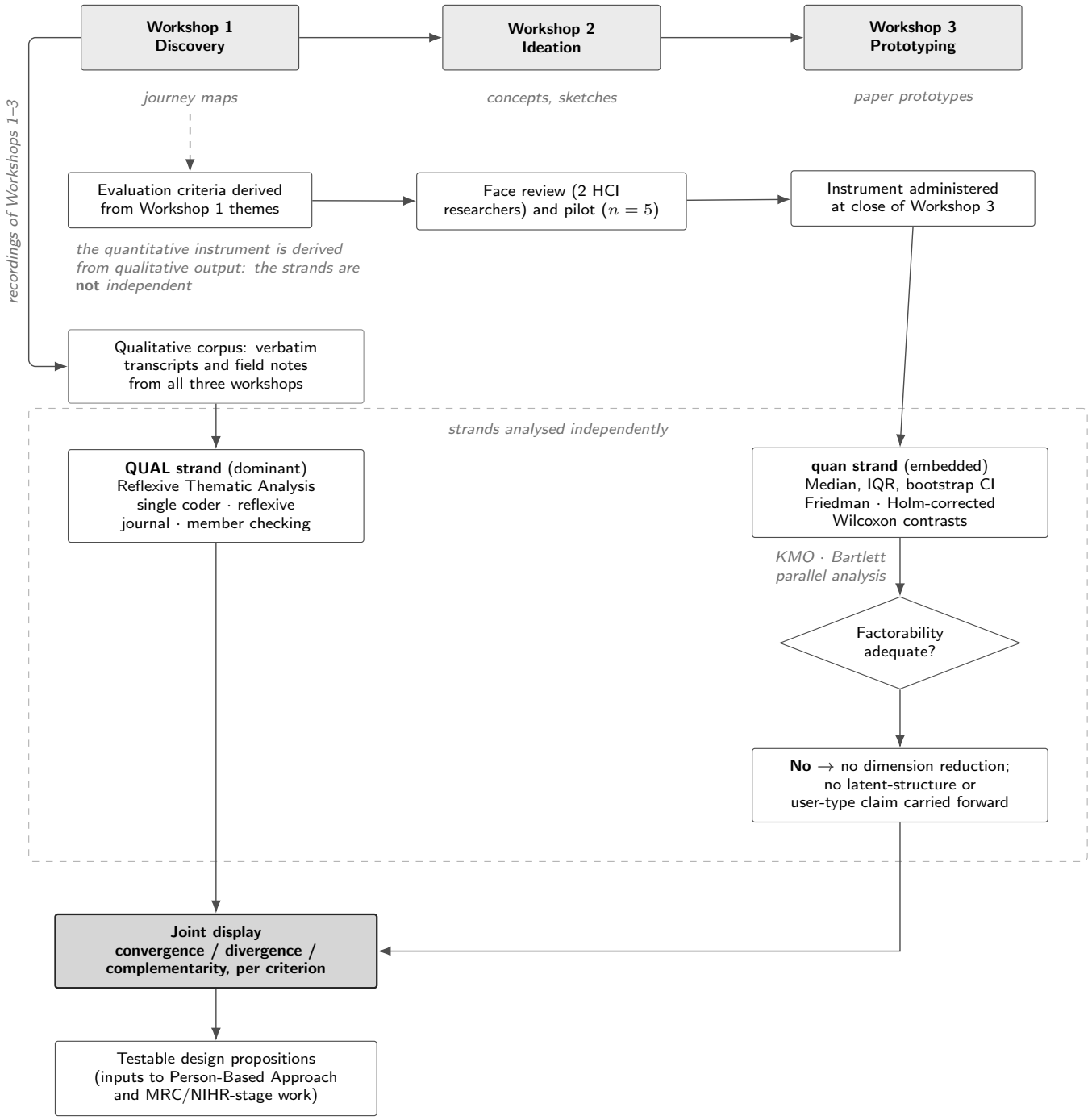


Figure 1. The structured mixed-methods workflow as implemented in this study. Solid arrows denote sequence; the dashed arrow denotes the derivation of the quantitative instrument from qualitative output; the dashed box encloses the stages at which the two strands are analysed independently; the light-shaded blocks are the three workshops; and the dark-shaded block marks the single point of formal integration.

Participation was incentivised with credit towards participants' degree programmes. An alternative task of equivalent workload was available to any student who preferred to earn the same credit without taking part in the study, and withdrawal from the study at any point did not affect entitlement to that credit. No member of the facilitation team

held a teaching or assessment relationship with any participant. Participants provided written informed consent prior to engagement and were assured of the anonymity and confidentiality of their data. The gender question offered non-binary and prefer-not-to-say response options; no participant selected either. The demographic characteristics of the participants are summarised in Table 1.

Table 1. Participant demographic characteristics (N = 45).

Characteristic	Category	Value (N/%)
Age (years)	Mean (SD)	22.1 (1.8)
	Range	19–25
Gender	Male	28 (62.2%)
	Female	17 (37.8%)
	Non-binary/prefer not to say	0 (0%)
Programme of study	Engineering	12 (26.7%)
	Architecture	12 (26.7%)
	Design	11 (24.4%)
	Ergonomics	10 (22.2%)

3.3. Data Collection Procedures and Instruments

3.3.1. Qualitative Stream: Participatory Co-Design Workshops

Participants were organised into nine multidisciplinary groups of five and took part in a series of three two-hour co-design workshops facilitated by a team of academic and industry researchers. Each group worked together across all three sessions and, in Workshop 3, built and then evaluated its own prototype, so the study yielded nine prototypes rather than one. Group membership is recoverable for every participant because participant identifiers were allocated group by group at the time of the workshops: S-01 to S-05 constituted Group 1 and so on to S-41 to S-45 as Group 9. A group identifier is included in the deposited rating matrix (Supplementary Materials, S2), and the consequences of this clustering for the quantitative analysis are treated in Section 3.4.1. All 45 participants attended all three sessions; complete retention is attributable to the scheduling of the workshops within timetabled hours and to the credit arrangement described above. The workshops were designed to be collaborative and generative, moving iteratively through stages of discovery, ideation, and prototyping [11]. Each workshop followed a structured protocol:

- **Introduction and insight playback (15 min).** Each session began with a summary of the key findings and design decisions from the previous workshop to ensure continuity and build a shared understanding. This also served as a form of member checking.
- **Generative activities (60 min).** The core of each workshop involved a primary creative task:
  - *Workshop 1 (Discovery).* User journey mapping. Participants mapped potential user journeys for a student seeking mental health support, identifying key pain points, opportunities, and emotional states. This activity grounded the design process in imagined lived experience and produced the thematic material from which the eleven evaluation criteria were subsequently derived.
  - *Workshop 2 (Ideation).* Brainstorming and concept sketching. Building on the journey maps, participants engaged in rapid idea generation to produce a wide range of candidate features and solutions. These ideas were clustered thematically and developed into low-fidelity paper sketches.
  - *Workshop 3 (Prototyping and feedback).* Interactive prototyping and think-aloud. Each group used its own sketches to create a low-fidelity, interactive paper prototype of the AMGR-I platform's core features, then engaged in a think-aloud

protocol, role-playing as users to test that prototype and provide real-time feedback on usability and clarity.

- **Reflection and consolidation (45 min).** The final part of each workshop was dedicated to a group discussion in which key insights were consolidated and priorities for the next design iteration were collectively decided.

Because the workshops required participants to engage with the experience of seeking mental health support, safeguarding provisions were in place throughout. At the start of each session, participants were told verbally that they could pause, step out, or withdraw at any point without giving a reason and without any consequence for their course credit, and that they were not required to disclose personal experience at any stage. Facilitators remained available during and after each session, and participants who wished to discuss anything arising from the session were directed to the university's student support and counselling services. These provisions were verbal and informal; no written distress protocol was in place, and this is acknowledged as a limitation. No participant withdrew from a session and no distress-related incident was recorded.

All workshops were audio-recorded and transcribed verbatim. Facilitators also took detailed field notes and collected all artefacts produced by the participants.

3.3.2. Quantitative Stream: Prototype Evaluation Instrument

At the conclusion of the final workshop, all 45 participants completed a structured, researcher-developed instrument in which they rated the low-fidelity AMGR-I prototype their own group had built and tested against eleven design criteria, each on a ten-point scale. The criteria were interactivity, gamification, immersiveness, usability, accessibility, personalisation and customisation, content quality, visual appeal, user experience, responsiveness, and data privacy and security. The instrument therefore measured participants' evaluation of the prototype against each criterion, not the general importance they attached to that criterion in the abstract.

We state this explicitly because the earlier account of this study described the instrument as measuring the importance of each criterion. That description was incorrect, and correcting it changes the interpretation of the results reported in Section 4.2.

Instrument development and pre-testing. The criteria were derived from the themes generated in Workshop 1 and cross-checked against recurring evaluation dimensions in the human–computer interaction (HCI) literature. To support clarity and consistency, the instrument underwent face review by two senior HCI researchers not otherwise involved in the study, and was pilot-tested with a separate group of five students to assess wording clarity and comprehension, resulting in minor refinements prior to administration. No psychometric validation was undertaken: no internal consistency coefficient was calculated, and no construct validity assessment or item analysis was performed. Moreover, the instrument should be understood as a purpose-built formative evaluation device rather than a validated scale.

Availability of the instrument. The original instrument file was not retained in a recoverable form. The account given above is provided by the first author, who designed and administered it, and is corroborated by the contemporaneous account of the study given in [23], which likewise describes the ratings as an evaluation of the prototype. The non-retention of the instrument is a documentation failure on the authors' part, it prevents independent verification of the item wording, and it is recorded as a limitation in Section 5.4.

3.4. Mixed-Methods Data Analysis Strategy

The qualitative and quantitative data were analysed independently before being integrated.

3.4.1. Quantitative Data Analysis

All quantitative analyses were conducted in Python 3.11 using NumPy 2.4.4, SciPy 1.17.1, pandas 3.0.2, factor_analyzer 0.5.1 and Matplotlib 3.10.9. The complete analysis script, which reproduces every statistic and figure reported in this paper from the raw rating matrix, is provided in the Supplementary Materials, S2.

For each criterion, we report the median, interquartile range, a bias-corrected bootstrap 95% confidence interval for the median based on 10,000 resamples, the mode, the observed range, and the mean and standard deviation. Because the same participants rated all eleven criteria, differences across criteria were tested with the Friedman test, with Kendall's W as the effect size. All 55 pairwise contrasts were then examined using Wilcoxon's signed-rank tests with Holm's correction applied across the full family of comparisons, and the rank-biserial correlation r reported as the effect size for each contrast.

Because each group rated the prototype it had itself built, the 45 rating vectors are clustered within nine artefacts, and the procedures described above treat as independent blocks observations that are not fully independent. Three further analyses were therefore conducted. First, the intraclass correlation coefficient $\mathrm{ICC}(1)$ was estimated for each criterion from a one-way random-effects decomposition, and the implied design effect computed at a group size of five. Second, the omnibus test was repeated on the nine group-mean profiles, which absorbs the clustering entirely at the cost of reducing the number of blocks to nine. Third, each of the 55 pairwise contrasts was re-tested using a linear mixed model fitted to the paired difference scores with a random intercept for group, with Holm's correction applied across the same family of 55; this is the primary clustered analysis and is reported alongside the participant-level results throughout. A further set of contrasts using cluster-robust standard errors referred to $t(8)$ is reported in the Supplementary Materials, S2, as a conservative bound, with the caveat that cluster-robust inference is known to be unreliable with as few as nine clusters. An exhaustive cluster sign-flip permutation test was also computed and is reported there uncorrected only, because with nine clusters, its smallest attainable p-value is $1/512 = 0.002$ and no contrast can survive Holm's correction across 55 comparisons at that resolution; the same resolution limit applies to a signed-rank test on nine group means, and neither is treated as evidence about the contrasts.

The factorability of the eleven criteria was assessed before any dimension-reduction procedure was attempted, using the Kaiser–Meyer–Olkin measure of sampling adequacy [52], Bartlett's test of sphericity, and Horn's parallel analysis based on 5000 permutations of the observed data [53]. Pearson's, Spearman's and polychoric correlation matrices were computed and compared, the polychoric matrix being the appropriate choice for ordinal responses. The outcome of these checks, reported in Section 4.4, determined that no component-extraction procedure was warranted.

3.4.2. Qualitative Data Analysis

The qualitative dataset comprises verbatim transcripts of the three workshops, including the Workshop 3 think-aloud sessions, together with facilitator field notes. No individual or small-group interviews were conducted at any stage, and no data source other than those named here contributed to the analysis. The preprint account lists interviews among the study's qualitative sources and attributes the quantitative data to usability tests; neither description is accurate, and the account given here is the correct one. The qualitative dataset was analysed using Reflexive Thematic Analysis [54]. This approach was chosen because

it acknowledges the researcher's active role in interpreting data and constructing meaning, which is appropriate for collaborative and design-oriented research contexts where researchers participate in the iterative development process [55]. The analysis followed the six-phase process: data familiarisation, generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report.

Coding and theme development were carried out by the first author working alone. This is a deliberate choice consistent with Reflexive Thematic Analysis, which treats the researcher's interpretative engagement as a resource rather than a source of bias to be controlled, and which explicitly does not require multiple coders, consensus coding or inter-rater reliability statistics [54]. In place of those procedures, the first author maintained a reflexive journal throughout, documenting analytical decisions and examining how his own position as both facilitator and analyst shaped the interpretation, including the bearing on that interpretation of his declared commercial interest in the platform concept, which is stated in the Conflicts of Interest statement and which gave him a stake in the artefacts he was analysing; themes were presented back to participants at the start of the two subsequent workshops for confirmation and refinement. Coding was performed manually on the transcripts and field notes; no qualitative analysis software was used. The preprint states that thematic analysis was carried out using software, which is not accurate. Reporting of the qualitative component follows the COREQ checklist [56]; a completed checklist is provided in the Supplementary Materials, S4, with any item we are unable to satisfy marked as not reported rather than omitted.

Consistent with this analytic framework, we do not report participant counts as an index of thematic importance. The earlier account of this study reported such counts; they have been removed, because Reflexive Thematic Analysis does not license them and because their derivation from group discussion, in which not every participant speaks on every topic, cannot be specified. Where we describe a topic as having been among the most frequently or forcefully discussed, as we do of data privacy, that is an analyst's characterisation of the sessions grounded in the transcripts and field notes, and not a count of participants or of utterances. We draw the distinction explicitly because the two are easily conflated and because the central finding rests on such a characterisation; a reader who declines to accept the characterisation should read the corresponding result as the artefact's low rating on a criterion participants raised unprompted and returned to across sessions, which is the weaker claim the retained material will carry on its own.

3.4.3. Data Integration

The quantitative results and qualitative themes were integrated through a joint display [51], presented in Section 4.5. For each of the eleven criteria, the display sets the quantitative result alongside the corresponding qualitative material and classifies the relationship between them as convergence, divergence or complementarity, before stating the resulting design implication. The joint display is the mechanism by which integration is performed in this workflow, and it is presented as a main table rather than an appendix for that reason.

3.5. Ensuring Methodological Rigour and Trustworthiness

The trustworthiness of the qualitative findings was addressed through the criteria set out by Lincoln and Guba [57], following the operationalisation described by Nowell and colleagues [58].

- *Credibility.* Credibility was addressed through triangulation, by comparing workshop findings with the prototype evaluation data—noting that the two strands are not independent, which limits what the comparison can establish, as set out in Section 5.4—and

through member checking: at the beginning of Workshops 2 and 3, a summary of themes from the previous session was presented to participants for confirmation and refinement.

- *Transferability.* The findings are not statistically generalisable. Transferability was addressed through thick description of the research context, participant characteristics and study procedures, allowing for readers to judge applicability to related design settings.
- *Dependability and confirmability.* These were addressed by maintaining an audit trail comprising verbatim transcripts, coded data, and a reflexive journal documenting analytical decisions, all of which are retained. The participant-generated artefacts were collected during the sessions but were not retained (Section 3.6), and the audit trail accordingly does not include them; the confirmability of the analysis rests on the transcripts, the coded data and the journal alone.

3.6. The AMGR-I Prototype

AMGR-I is the name of the prototype platform developed in this study; it is not an initialism, and the suffix denotes the first pilot iteration. The concept was for an integrated digital mental health platform for a university student population, organised around four sub-systems identified by participants in Workshops 1 and 2 and elaborated in Workshop 3: a social and community sub-system supporting peer connection; a health sub-system mediating contact with care providers; an e-learning sub-system delivering psychoeducational content; and a cloud sub-system handling storage and data sharing. Figure 2 shows the conceptual structure that emerged from this process. The preprint account of this study presents the same four-part structure in introductory material about the platform rather than as a study output. That preprint was written five months after the workshops, by which time the structure the sessions had produced was simply the platform's architecture; its placement there reflects the order of writing, not the order of events. The platform was conceived as a signposting, psychoeducation and peer-support environment; it was not designed to make or recommend treatment decisions, and no clinical decision support was prototyped or evaluated. Figure 2 retains therapy, medication and insurance providers as nodes because the structure locates the platform in relation to those external actors; the nodes denote structural relationships, principally signposting and the exchange of information with a provider the user has chosen, and not functionality that was prototyped or rated. No node in the figure corresponds to a clinical decision-support feature.

In Workshop 3, each of the nine participant groups constructed its own low-fidelity interactive paper prototype of the platform's core screens and flows, working from the sketches produced in Workshop 2. Three design directions were explored across the nine groups—minimalist, illustrative and immersive—of which the immersive direction attracted the most favourable comment. It is these nine group-built paper prototypes, and no running software, that participants evaluated against the eleven criteria, each group rating the artefact it had itself produced. The physical prototypes and any photographs of them were not retained, and we have not reproduced them, since a reconstruction made by the authors would show something other than the participant-made artefacts that were rated. Table 2 instead records what was prototyped. It rests on the first author's account of the sessions, corroborated by the contemporaneous preprint [23] as to the four sub-systems and the three design directions. The preprint does not corroborate the medium, for the reason set out immediately below, and the inventory should not be read as resting on it in that respect.

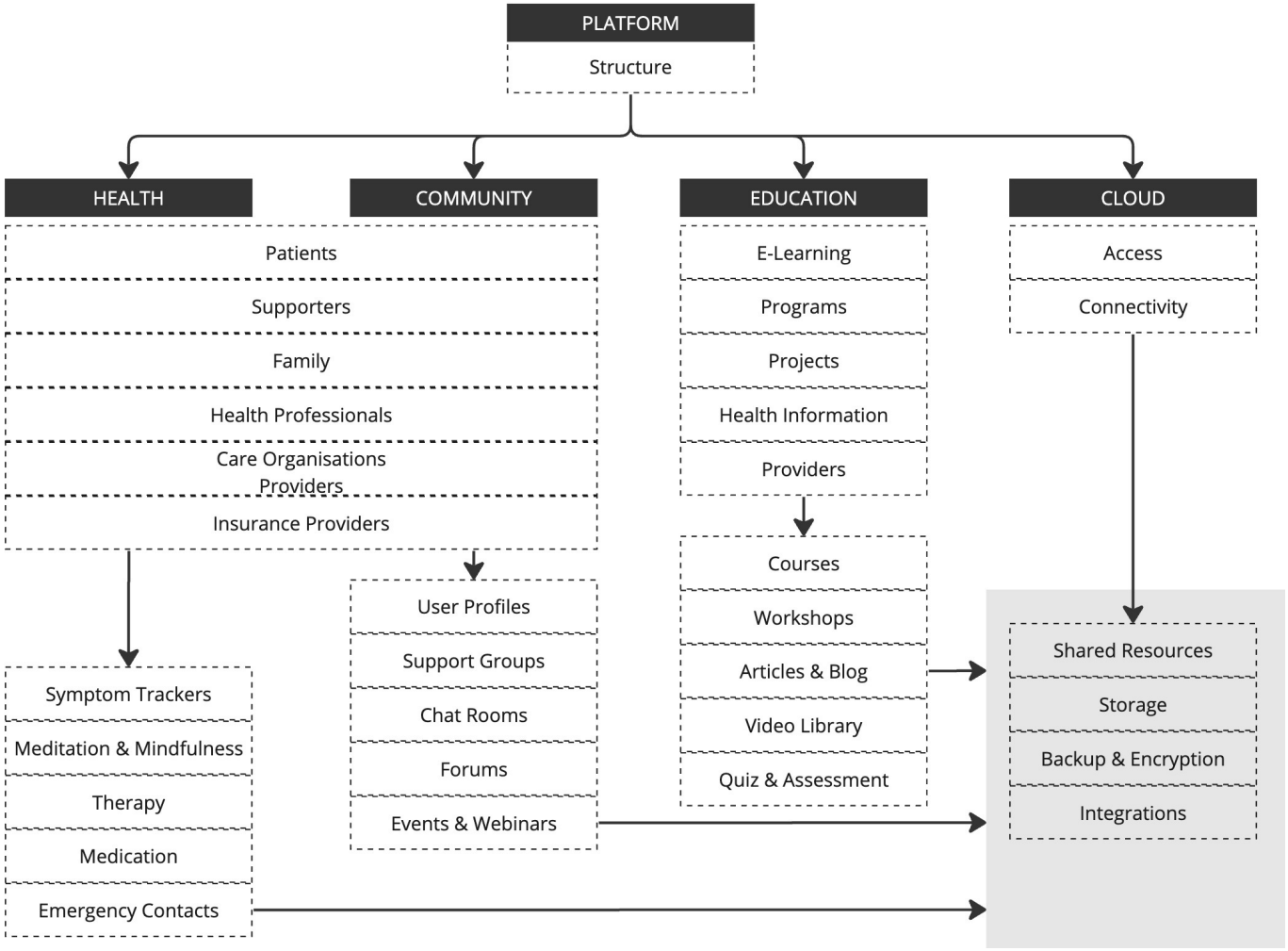


Figure 2. The conceptual structure of the AMGR-I platform, as developed through the co-design workshops. Dark header bars denote the four sub-systems; dashed boxes denote the actors and components identified within them. Arrows from the platform node denote its decomposition into sub-systems; arrows between groups of components denote the order of elaboration; and arrows into the shaded region denote the flow of content and user data to the storage, backup and integration functions of the cloud sub-system.

Reconciliation with the preprint account of prototype fidelity. The preprint describes, in its introductory account of the platform rather than in its methods, a design progression from paper sketches through low- and mid-fidelity digital wireframes to high-fidelity interactive prototypes built in Figma, and states that high-fidelity interactive prototypes were employed during an evaluation phase. That passage describes the authors' own development of the AMGR-I concept, which ran separately from the study reported here and continued after it. It does not describe what the participants in this study evaluated. The artefacts rated against the eleven criteria were the nine paper prototypes built by the participant groups in Workshop 3. No digital, interactive or running prototype was placed before participants at any point in the three workshops, and no rating reported in this paper refers to one. We state this explicitly because the preprint's placement of that passage makes it easy to read as a description of this study's method, and because the two accounts have been read as contradictory. One principle governs our use of the preprint throughout, and we state it rather than leave it to be inferred: it is a contemporaneous record and we rely on it where it corroborates the authors' own records, as it does on the construct the instrument measured; where it conflicts with those records on matters of procedure, the account given

in this paper is the corrected one and we identify each such conflict individually rather than setting the preprint aside wholesale. The high-fidelity files themselves were produced in an institutional account to which the first author no longer has access and cannot now be retrieved. We record that for completeness, since the preprint names the tool and such files are ordinarily recoverable, and not as a limitation of the present study: those artefacts were not evaluated by participants, and no result reported here depends on them.

Table 2. What was prototyped and evaluated. The physical prototypes were not retained (see Section 3.6).

Element	What Was Prototyped or Explored
Social and community sub-system	Community and social-networking functions supporting peer connection among students
Health sub-system	Interaction with healthcare providers; signposting to care rather than delivery of it
E-learning sub-system	Delivery of educational and psychoeducational content
Cloud sub-system	Storage and sharing of user data across the platform
Design directions	Three directions explored in parallel—minimalist, illustrative and immersive—with the immersive direction attracting the most favourable comment
Medium	Nine low-fidelity interactive paper prototypes of core screens and flows, one per participant group, built from that group's Workshop 2 sketches; no digital or running software at any stage
Testing procedure	Think-aloud protocol, with participants role-playing as users
Explicitly out of scope	No clinical decision support; no treatment recommendation; no automated diagnosis

4. Results

The integrated analysis identified three themes concerning how participants evaluated the AMGR-I prototypes and what they wanted from a digital mental health platform. Table 3 sets out the themes and sub-themes with definitions and illustrative extracts.

Table 3. Themes, sub-themes and illustrative extracts. Verbatim extracts appear in italics within quotation marks. All extracts in this table are verbatim; the paraphrased entries in the earlier version of this table have been replaced with the participants' own wording, and ellipses mark omitted material within a quotation. The workshop (W) is given where it was recorded. In one extract, a detail that could have identified a participant within the cohort has been generalised; the generalisation is marked in the text where it occurs.

Theme	Sub-Theme	Definition	Illustrative Extract
1. Legibility under duress	Simplicity and clarity	The platform must be operable by someone whose cognitive resources are depleted; interface complexity is read as a barrier to help, not a feature.	*"I don't want to have to learn how to use an app when I'm feeling stressed. It needs to be completely intuitive, almost invisible"* [P16, W3]
	Credibility of content	Information quality is treated as a precondition of use rather than one attribute among others.	*"Does this content we depend on come from professionals or from other people in the community going through the same problems as us? One is reliable and the other might be misleading information"* [P12, W3]

Table 3. *Cont.*

Theme	Sub-Theme	Definition	Illustrative Extract
2. Security as a spoken absolute and an evaluated shortfall	Data protection as a threshold condition	Privacy is articulated as non-negotiable and prior to any other consideration.	*"This is the most sensitive data imaginable. There needs to be zero doubt that it's secure; otherwise, no one will use it, no matter how good the features are"* [P41]
	Trust as something shown, not stated	Participants distinguished between a platform being secure and a platform demonstrating that it is secure.	*"I am really concerned about my identity being out there … who sees what after a digital verification and is this information out there permanent? Who has access to it and what kind of privacy … can we expect after what we share?"* [P17, W3]
3. Innovation conditional on therapeutic integrity	Personalisation with human oversight	Adaptive features are welcomed only where a human remains accountable for consequential judgements.	*"I like the idea of it learning what I need, but I'd be worried about a computer making therapy decisions without a real person checking it"* [P8]
	Gamification as effort and as avoidance	Game mechanics are read as carrying a cost in attention and persistence that a user in distress may not have, and as risking displacement of the problem rather than engagement with it. A distinct concern that such mechanics trivialise therapeutic purpose is also recorded.	*"That is a playful distraction but might also at the same time drain me if am already tired and worried about other things in my life"* [P11]; *"Would that be a good thing or a temporary distraction and escape from life and reality rather than dealing with my problems?"* [P37]

An illustration of the coding trajectory may clarify the analytic process. Codes such as *cannot think straight when stressed*, *too many steps*, and *wants it to disappear into the background* were grouped into the sub-theme *simplicity and clarity*, which together with *credibility of content* formed the theme *legibility under duress*. The themes are analytic constructions, not groupings of the survey criteria, and they cut across those criteria rather than mapping onto them one to one.

4.1. Evaluation of the Prototypes Across the Eleven Criteria

Table 4 reports the distribution of ratings for each criterion. Figure 3 shows the same distributions, with individual ratings overlaid so that dispersion and clustering are visible rather than summarised.

Ratings differed systematically across the eleven criteria (Friedman $\chi^2(10) = 225.73$, $p < 0.001$, Kendall's $W = 0.50$, indicating a moderate degree of agreement among participants in how they ranked the criteria). Of the 55 pairwise contrasts, 41 remained significant after Holm's correction; the full matrix of contrasts with effect sizes is provided in the Supplementary Materials, S2.

Because each group rated the artefact it had itself built, these participant-level tests were repeated with the clustering modelled, as set out in Section 3.4.1. Intraclass correlations by criterion range from 0 to 0.31 with a median of 0.09, five of the eleven exceeding 0.10, which at a group size of five implies a design effect of 1.35 at the median and 2.26 at the maximum. The omnibus difference across criteria survives when the clustering is

absorbed entirely by analysing the nine group-mean profiles (Friedman $\chi^2(10) = 71.29$, $p < 0.001$, Kendall's $W = 0.79$). Under the mixed model with a random intercept for group, 40 of the 55 contrasts remain significant after Holm's correction against 41 at participant level: two are lost, usability against content quality and responsiveness against data privacy and security, and one is gained, immersiveness against user experience. Under the deliberately conservative cluster-robust treatment reported in the Supplementary Materials, S2, 24 remain significant. The ordering of the criteria is therefore robust to the clustering; the number of pairs that can be distinguished is not. Mixed-model results are reported below, with participant-level results given alongside.

Table 4. Distribution of prototype evaluation ratings by criterion (N = 45), ordered by median. Several bootstrap intervals for the median are degenerate (usability, accessibility and visual appeal 8.0–8.0; interactivity 7.0–7.0; data privacy and security 6.0–6.0). This follows from a discrete bounded ten-point scale with heavy ties: in every resample, the median falls on the same scale point, so the interval collapses to it. These intervals should be read as indicating that the median is stable under resampling, not as evidence of precision finer than one scale point.

Criterion	Mdn	IQR	95% CI	Mode	Range	Mean	SD
Content Quality	9	8–10	9.0–10.0	10	6–10	8.89	1.25
Usability	8	7–9	8.0–8.0	8	5–10	7.91	1.06
Accessibility	8	7–8	8.0–8.0	8	5–10	7.87	1.08
Visual Appeal	8	7–9	8.0–8.0	8	6–10	7.84	1.17
Personalisation and Customisation	7	7–8	7.0–8.0	8	6–9	7.29	0.94
Interactivity	7	6–8	7.0–7.0	7	4–9	6.98	1.14
Responsiveness	7	6–7	6.0–7.0	7	4–9	6.69	0.97
User Experience	7	6–7	6.0–7.0	7	5–9	6.67	0.95
Immersiveness	6	5–7	5.0–6.0	6	4–8	5.98	1.14
Data Privacy and Security	6	5–7	6.0–6.0	6	3–8	5.89	1.05
Gamification	5	5–6	5.0–6.0	6	3–7	5.20	0.92

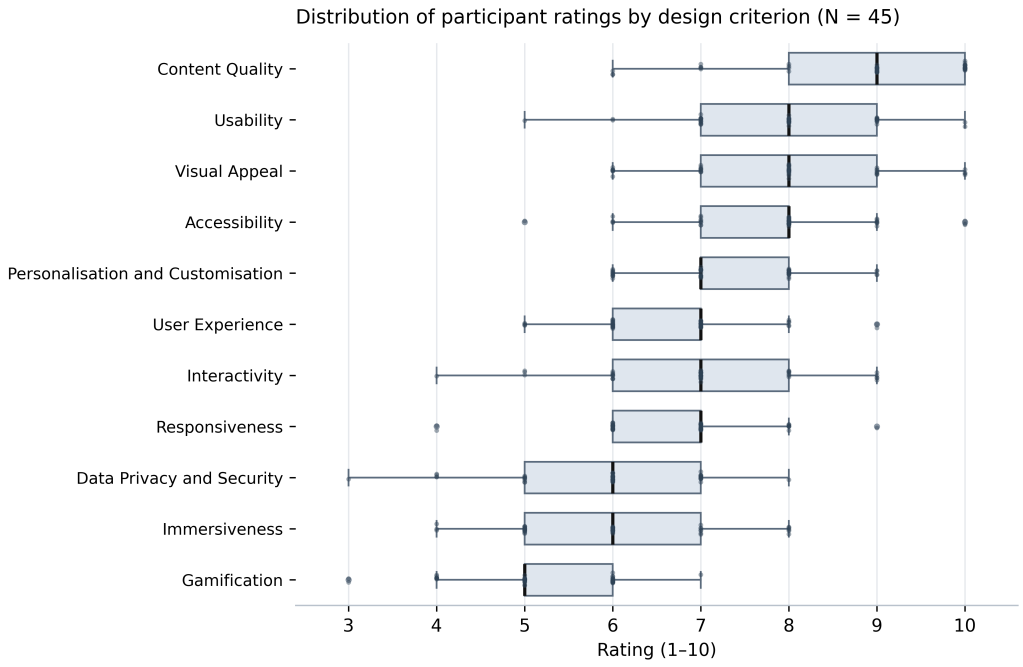


Figure 3. Distribution of participant ratings by design criterion (N = 45). Boxes show the interquartile range and median; individual ratings are overlaid with horizontal jitter.

Two of the ordering claims made in the earlier account of this study are broadly supported, and one requires correction. Content quality was the highest-rated criterion, significantly higher than nine of the other ten under the mixed model and than all ten at participant level; once clustering is modelled, it is not distinguishable from usability. Gamification was the lowest-rated criterion, significantly lower than each of the other ten under both treatments. Data privacy and security was rated significantly lower than seven of the ten remaining criteria under the mixed model and eight at participant level; it is not distinguishable from immersiveness under either treatment, nor from responsiveness once clustering is modelled; and it was rated significantly *higher* than gamification, which has the lower mean of the two, 5.20 against 5.89. The earlier version of this section stated that data privacy and security was rated lower than nine of the ten remaining criteria. That was an error. The analysis script reports the number of criteria from which privacy differs, and nine is that number, but one of the nine is gamification, where privacy is the higher of the pair; rendering *differs from* as *lower than* produced a claim that cannot hold. It is accurate to describe data privacy and security as among the lowest-rated criteria, and inaccurate to describe it either as having been prioritised or as the lowest-rated.

4.2. Security as a Spoken Absolute and an Evaluated Shortfall

The most informative result of this study concerns data privacy and security. It is visible only because the two strands were analysed separately and then set against one another.

In the workshops, data protection was among the most frequently and forcefully discussed concerns. Participants raised it unprompted, returned to it across sessions, and framed it as a threshold condition rather than a desirable feature: *"This is the most sensitive data imaginable. There needs to be zero doubt that it's secure; otherwise, no one will use it, no matter how good the features are"* [P41]. Concern extended beyond encryption to what becomes of disclosed information once it has been given: *"I am really concerned about my identity being out there. Like, who sees what after a digital verification and is this information out there permanent? Who has access to it and what kind of privacy about this and our personal situations can we expect after what we share? We would be sharing a lot"* [P17, W3]. Discussion consistently emphasised end-to-end encryption, transparent data-usage policies, and the persistence and visibility of what participants would disclose.

Yet when the same participants evaluated the prototype they had built and tested, data privacy and security received a median of 6 (IQR 5–7), placing it among the lowest-rated criteria: significantly below seven of the other ten under the mixed model and eight at participant level, indistinguishable from immersiveness, and above only gamification.

Because the instrument asked participants to evaluate the prototype rather than to rate the abstract importance of each criterion, this pattern is not a dissociation between attitude and evaluation. It is convergence. Participants who regarded protection as indispensable rated their own artefact poorly on precisely that attribute, which is what one would predict, and the qualitative material therefore explains the quantitative result rather than standing in tension with it. Privacy was salient in what participants said and absent from what they made. The earlier version of this section presented the finding as a divergence and set it alongside the privacy paradox literature. Under the corrected construct, both moves were wrong and both are withdrawn: the privacy paradox concerns the gap between stated concern and privacy-protective behaviour, and rating an artefact's provision is neither a statement of concern nor a protective behaviour. What remains is narrower and more useful. The requirement a design team discussed most was the one its artefacts served worst.

That reading is sharpened by who the design team was. Each group rated the prototype it had itself built, so the criterion these design-literate participants named as a threshold condition is the one they did not build into their own artefacts. The omission is theirs, and it

was observed rather than reported. That is a stronger result than a discrepancy between what a group says and how it judges somebody else's work, and it is reported here as a finding in its own right.

A second explanation of the low rating sits alongside the first, and is arguably the stronger of the two: a limit of the medium. A paper prototype cannot demonstrate encryption, a retention policy or an access-control model, so there was little for participants to rate favourably even where the design had considered them. We make the same argument for responsiveness and interactivity in Section 5.4, and it applies with greater force to security, which is less representable in that medium than either. This does not weaken the design lesson but specifies it: a protection that cannot be made perceptible in a low-fidelity artefact is also one a user cannot verify in a finished product. Both of these readings, the omission and the limit of the medium, lead to the same conclusion about perceptibility, and it is on that conclusion that the design implication rests.

One alternative account would undercut it, and cannot be excluded. Three sessions in which privacy was repeatedly raised may have made participants more exacting in evaluating their own artefact's treatment of it than they would otherwise have been, so that the rating reflects a raised standard of judgement rather than a shortfall in the artefact. An earlier version of this passage stated this alternative in terms of the discursive prominence of privacy relative to the weight participants individually assigned it; that formulation reintroduced the importance construct which Section 3.3.2 disclaims, and it has been restated above in terms of the evaluation itself. The design of this study cannot distinguish a raised standard of judgement from a genuine shortfall, and we do not claim to.

4.3. Conditional Acceptance of Advanced Features

Participants' relationship to advanced platform functionality was consistently conditional. Personalisation attracted substantial qualitative support, with participants describing tailored content and recommendations as desirable, but that support was routinely qualified by a demand for human oversight and a wariness of automated judgement: *"I like the idea of it learning what I need, but I'd be worried about a computer making therapy decisions without a real person checking it"* [P8]. In the evaluation data, personalisation and customisation sat in the middle of the distribution (median 7, IQR 7–8).

Gamification attracted the clearest negative response in both strands. It received the lowest rating of any criterion (median 5, IQR 5–6) and was significantly lower than every other criterion after correction. The qualitative material bears directly on this result, and it is more specific than the criterion's low rating alone suggests. Participants did not object to game mechanics as undignified so much as doubt that they could afford them. The recurring concern was the cost in attention and persistence that such mechanics impose on someone already depleted: *"Gamification is engaging but not sure how much of this I can sustain"* [P6]. Closely related was a concern about displacement, that a mechanic designed to engage might instead substitute for the problem it was meant to help with (*"Would that be a good thing or a temporary distraction and escape from life and reality rather than dealing with my problems?"* [P37]). A countervailing view was expressed, grounding acceptance in prior familiarity rather than in design: *"I have gamed as a child growing up in [another country]. It almost identifies as a culture and way of life that under stress it might be a great way to cope"* [P8]; the country named by the participant has been generalised, since in a cohort of this size and institutional concentration it could have identified them. A distinct concern that game mechanics would trivialise the platform's therapeutic purpose is recorded in the earlier account of this study. We retain it as a separate strand because that account is contemporaneous, but no extract in the retained transcripts carries that wording, and we state that rather than illustrating the strand with material that does not support it. We do

not characterise which position predominated. An earlier version of this section stated that more participants expressed concern than saw potential; that is a participant count used as an index of thematic weight, which Section 3.4.2 states we do not report, and it has been withdrawn. The two positions are reported as co-present, and the ordering of the criteria is left to the evaluation data, where gamification is unambiguously last. The tension the material records is not between engagement and disengagement but between engagement and what a user in distress can sustain.

4.4. Factorability of the Eleven Criteria

We had intended to examine whether the eleven criteria reduced to a smaller number of underlying dimensions, and the earlier account of this study reported a two-component solution with interpretative labels. On formal assessment, the data do not support any component-extraction procedure, and that reported solution is withdrawn.

The ratio of cases to variables is 4.1:1, below the 5:1 to 10:1 minima commonly recommended. The Kaiser–Meyer–Olkin measure of sampling adequacy is 0.42 overall, below the 0.50 threshold at which factoring is conventionally judged unacceptable [52]; ten of the eleven individual measures of sampling adequacy also fall below 0.50, ranging from 0.29 to 0.50. Bartlett's test of sphericity is non-significant ($\chi^2(55) = 54.88$, $p = 0.479$), meaning the correlation matrix cannot be distinguished from an identity matrix. Horn's parallel analysis retains no components: the first observed eigenvalue is 1.92 against a 95th-percentile simulated value of 2.16 (Figure 4) [53]. Eigenvalues computed from the polychoric matrix, which is the appropriate choice for ordinal data, give the same result. Kaiser's eigenvalue-greater-than-one rule would retain five components, which at this sample size is a well-documented over-extraction artefact rather than a competing result.

Taken together, these checks indicate that participants' ratings of the eleven criteria were close to mutually independent. There is no latent structure here to recover, and consequently no basis for describing participants as distributed along dimensions or as falling into preference types. The correlation matrices, eigenvalues, parallel-analysis thresholds and, for completeness, the rotated loading matrices that would have resulted are provided in the Supplementary Materials, S2. We note that those loadings would not have supported the labels previously assigned in any case: usability and data privacy, two of the four criteria cited as defining the first component, load on neither component above 0.40, and the communality of data privacy is 0.04. The withdrawn two-component solution accounted for 17.45% and 14.52% of variance, 31.97% cumulatively, and under an oblique promax rotation, the two components correlated at −0.05, so the oblique solution recovered nothing the orthogonal one had missed. A further point tells against the withdrawn labels more directly than the loadings alone. The first component was defined by content quality (−0.74), visual appeal (−0.68) and gamification (−0.55), with accessibility marginal at −0.43; it was not defined by the foundational usability and trust criteria the earlier account named. The labels were therefore not merely unsupported at this sample size; they did not describe the solution to which they were attached, which makes the negative result the more instructive.

This is a negative result, and we report it because it carries a methodological lesson for the workflow itself: a formative sample large enough to generate rich design material is not thereby large enough to support latent-variable claims about it, and the temptation to extract structure from such data should be resisted at the adequacy-testing stage rather than at the interpretation stage.

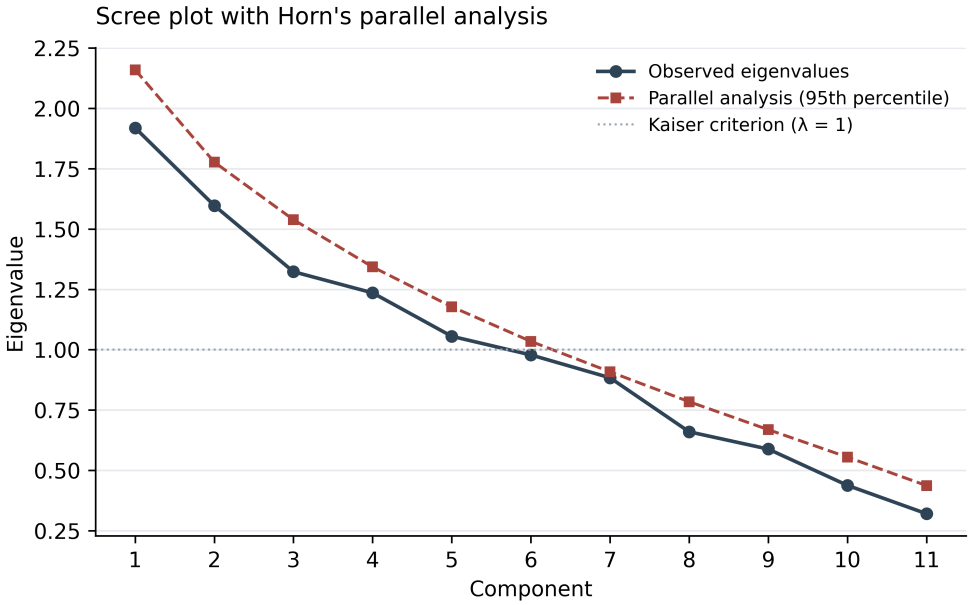


Figure 4. Scree plot of the eleven eigenvalues of the Pearson correlation matrix against the 95th-percentile eigenvalues from Horn's parallel analysis (5000 permutations). No observed eigenvalue exceeds its parallel-analysis threshold, so no component is retained. The dotted line marks Kaiser's criterion ($\lambda = 1$), which would retain five components and over-extracts at this sample size.

4.5. Integration of the Two Strands

Table 5 presents the joint display through which the strands were integrated.

Table 5. Joint display: Integration of evaluation ratings and qualitative themes.

Criterion	Mdn [IQR]	Corresponding Qualitative Material	Relationship	Design Implication
Content Quality	9 [8–10]	Credibility of information treated as a precondition of use; concern about professional versus community-generated content and misleading information [P12]	Convergence	Invest in editorial and clinical quality assurance before feature expansion
Usability	8 [7–9]	"Almost invisible" interaction demanded for use under distress [P16]	Convergence	Optimise for depleted cognitive resources, not for feature discoverability
Accessibility	8 [7–8]	Inclusive design raised as a condition of reach	Convergence	Treat accessibility conformance as scope, not enhancement
Visual Appeal	8 [7–9]	Immersive design direction preferred among the three explored	Complementarity	Visual register signals seriousness; align it with therapeutic tone
Personalisation and Customisation	7 [7–8]	Endorsed, but conditional on human oversight and explicability [P8]	Complementarity	Ship personalisation with visible rationale and user control; avoid opaque adaptation
Interactivity	7 [6–8]	No distinct qualitative material identified	Not evidenced	No implication claimed
Responsiveness	7 [6–7]	No distinct qualitative material identified	Not evidenced	No implication claimed

Table 5. *Cont.*

Criterion	Mdn [IQR]	Corresponding Qualitative Material	Relationship	Design Implication
User Experience	7 [6–7]	Overlaps with legibility-under-duress material rather than forming a distinct strand	Complementarity	Treat as composite; not independently actionable
Immersiveness	6 [5–7]	Interest expressed alongside concern about distraction from purpose	Complementarity	Use immersion to support focus, not to extend session time
Data Privacy and Security	6 [5–7]	Among the most forcefully discussed concerns; framed as a threshold condition [P41]; identity, verification and access management raised [P17]	Convergence	Make protection perceptible at the interface; a policy that is not shown is not experienced as protection
Gamification	5 [5–6]	Interest in game mechanics as an engagement device, expressed alongside concern about the attention and persistence they demand and about displacement of the underlying problem [P6, P37]	Complementarity	Align any game mechanic with a therapeutic goal, or omit it

Nine of the eleven criteria show convergence or complementarity between the strands, and none shows divergence. Data privacy and security was classified as a divergence in the earlier version of this display; under the corrected construct, it is a convergence, for the reasons set out in Section 4.2, and it is that convergence which produced the study's most actionable finding. Gamification moves in the opposite direction, from convergence to complementarity, because the qualitative material records interest as well as concern and therefore does not simply agree with the rating. Two criteria have no distinct qualitative material behind them; we record that absence rather than constructing an implication without evidence.

5. Discussion

This study set out to develop and evaluate a structured mixed-methods workflow for the formative stages of digital mental health platform development. The discussion interprets the findings, states them as testable propositions rather than as established principles, sets out the methodological contribution, and addresses the study's limitations, which are substantial.

5.1. Interpretation of the Findings

The evaluation data place foundational attributes—content quality, usability, accessibility, and visual appeal—above engagement-oriented ones, and place gamification last by a clear and statistically supported margin. Read alongside the workshop material, the pattern suggests that for this group of design-literate evaluators, a digital mental health platform earns the right to be engaging only after it has become legible and credible.

The data privacy result is the finding with the most direct design consequence. Participants who spoke about security as an absolute precondition rated it among the weakest attributes of the artefacts they had themselves built. The two readings the evidence supports, an omission in the artefact and a limit of what a paper prototype can demonstrate, converge on a single lesson: security which is architecturally present but not perceptible is, experientially, not present at all, because trust is produced at the interface through what a user can see and verify, and not solely in the data layer. We state that lesson with the

confidence the convergence of the two readings warrants, and no more, since the study cannot exclude the alternative set out in Section 4.2, namely that three sessions of discussion raised the standard against which participants judged their own work. It is a claim about how to build, and it is testable.

The conditional acceptance of personalisation points the same way. Participants did not reject adaptive functionality; they rejected unaccountable adaptive functionality. The relevant design variable is not how much the system adapts but how visible and contestable its adaptation is.

5.2. From Findings to Testable Design Propositions

We state the following as hypotheses arising from the study, to be tested with clinical and non-design populations, rather than as principles established by it. The distinction matters: these propositions rest on 45 design students evaluating a paper prototype, and that evidence base supports conjecture, not prescription.

- *Proposition 1: Foundational adequacy precedes engagement.* Resource allocated in early-stage design to usability, content credibility and perceptible security will yield greater sustained engagement than equivalent resource allocated to novel features. *Test:* A factorial comparison of development effort against retention in a deployed intervention.
- *Proposition 2: Personalisation is accepted in proportion to its explicability and controllability.* Adaptive features accompanied by visible rationale and user override will be accepted where functionally identical opaque features are not. *Test:* An experimental manipulation of explanation and control at constant underlying functionality.
- *Proposition 3: Gamification is accepted in proportion to its sustainability under depletion and its alignment with therapeutic purpose.* Game mechanics that make low demands on attention and persistence, and that represent progress through therapeutic content, will be received more favourably than achievement or competition mechanics of equivalent intensity, and the difference will be larger among users with a higher symptom burden. *Test:* A comparison of mechanic types on perceived therapeutic credibility, on engagement, and on sustained use, stratified by symptom severity.

5.3. Methodological Contribution

The contribution of this study is the workflow set out in Figure 1, together with an honest account of its limits. Its value lies in three features. First, it specifies where the quantitative instrument comes from: the criteria are generated by participants in the first workshop rather than imported from an external scale, which keeps the instrument grounded in the design context. Second, it specifies where integration happens: at a single explicit joint display, rather than diffusely across a narrative. Third, and demonstrated here by negative example, it places a formal factorability check before any structural claim about the quantitative data, so that a formative sample is not made to carry inferential weight it cannot bear.

We make no claim that this workflow improves on the Person-Based Approach or MRC/NIHR guidance; it operates earlier and at lower cost than either, and its output is properly an input to them. The intended user is a small design team at the concept stage with access to participants for a few sessions and no budget for a discrete choice experiment.

5.4. Limitations

The limitations of this study are consequential and constrain every finding reported above.

The sample. Participants were design, engineering, architecture and ergonomics students with no declared lived experience of mental ill-health. They are design-literate

formative evaluators, not users or proxies for users. Their judgements about what a person in distress would need are informed inferences, not reports. No finding here should be read as establishing the needs of people who would use a digital mental health platform, and the gender distribution (62.2% male) further limits transferability given documented gender differences in mental health help-seeking.

Sequential contamination. The instrument was administered after three sessions of group discussion on the same topics, and its criteria were derived from those discussions. The two strands are therefore not independent, which weakens the triangulation logic that would otherwise justify integration, and provides a plausible alternative account of the central finding, as set out in Section 4.2. In a future application of this workflow, administering the instrument before the final discussion, or to a non-overlapping group, would substantially strengthen it.

The instrument. The instrument was purpose-built and not psychometrically validated, and the original item file was not retained. The construct it measured is stated on the first author's account and corroborated by the contemporaneous preprint, but cannot be independently verified. This is a documentation failure and a real limitation on the reproducibility of the quantitative strand.

Absence of trade-off. Participants rated each criterion independently, which imposes no trade-off and produces the expected compression towards the upper part of the scale: ten of eleven criteria have a median of 6 or above. Ratings obtained this way indicate relative standing but do not establish priorities in the sense required for resource allocation. Methods that force trade-offs—best–worst scaling, discrete choice experiments, pairwise comparison, and point-budget allocation—are the appropriate instruments for that purpose [36,37], and we have accordingly tempered the language of prioritisation throughout.

No latent structure. As reported in Section 4.4, the data do not support dimension reduction, and no claim about underlying constructs or user types is made.

Documentation practice. Three classes of material from this study cannot now be produced: the instrument file, the participant-made paper prototypes, and the first author's own high-fidelity design file, which is held in an institutional account to which he no longer has access. These are separate failures with separate causes, but their number is itself informative. The study was conducted without a data-management plan, and instituting one is the single change most likely to improve a future application of this workflow. What does survive is what the reported findings rest on: the rating matrix and the complete analysis code, deposited in full in the Supplementary Materials, S2, and the workshop transcripts and field notes, which are retained.

Single coder. Coding and theme development were conducted by one researcher who was also a workshop facilitator. This is methodologically consistent with Reflexive Thematic Analysis and was mitigated by reflexive journalling and member checking, but it remains a single interpretative perspective.

Elapsed time. Data were collected in June 2023. Attitudes towards AI-driven personalisation in particular have developed considerably since, and the ratings for personalisation and related criteria should be read with that interval in mind.

Prototype fidelity. Participants evaluated low-fidelity paper prototypes, not working software. Criteria such as responsiveness and interactivity are only partially assessable in that medium, which may account for their undifferentiated mid-scale ratings, and data privacy and security is less assessable still, as discussed in Section 4.2.

Clustered ratings and self-evaluation. Each group rated the artefact it had itself built, so the ratings are clustered within nine prototypes and participants were evaluating their own work. The clustering is modelled in Section 4.1 and the substantive ordering is robust to it, but nine clusters is a small number: the mixed model is fitted to few groups, cluster-

robust inference is unreliable at this scale, and the number of distinguishable pairs varies appreciably with the treatment chosen. Self-evaluation is the more consequential of the two. Ratings of one's own artefact are not independent of the investment made in producing it, and the direction of any resulting bias is not predictable: a group may rate its own work generously, or may be more exacting about it than an outsider would be. A future application of this workflow should have groups evaluate artefacts other than their own, which would remove both problems at once.

5.5. Directions for Future Research

The immediate priority is to repeat the workflow with people who have lived experience of mental ill-health, and to test the three propositions above in that population. A second priority is instrument development: the eleven criteria would benefit from formal item development and, at an adequate sample size, a trade-off-based elicitation method. A third is longitudinal work, examining how evaluations of a deployed platform change with sustained use, and whether perceptible security measures affect retention as Proposition 1 predicts.

6. Conclusions

This study presented a structured mixed-methods workflow for the formative development of digital mental health platforms and applied it to the development of the AMGR-I prototype with 45 design-literate formative evaluators. The workflow combines participatory co-design with a complementary quantitative evaluation whose criteria are generated by participants themselves, and integrates the two through an explicit joint display.

Applied here, it produced one substantive finding and one methodological one. Substantively, participants evaluated their prototypes most favourably on content quality, usability and accessibility and least favourably on gamification, while data privacy and security—the concern they voiced most forcefully in discussion—was among the attributes they rated lowest in the artefacts they had themselves built. The readings the evidence supports converge on the lesson that protection which is not made perceptible is not experienced as protection, although the study cannot exclude the alternative that sustained discussion raised the standard against which participants judged their own work. Methodologically, the formal factorability check built into the workflow established that a sample of this size cannot support claims about latent structure, and prevented an unsupportable interpretation from being carried forward.

The design propositions offered here are conjectures for testing, not established principles, and the evidence supporting them comes from students evaluating a paper prototype rather than from people using a mental health intervention. What the paper offers practitioners is a procedure that is cheap enough to run at the concept stage, explicit about where its two strands meet, and candid about the inferences it cannot support.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/informatics13110178/s1, S1: The prototype evaluation instrument, describing its eleven criteria and administration procedure; S2: Data and reproducible analysis, comprising the complete analysis scripts, the raw rating matrix with group identifiers, descriptive statistics, the full matrix of Holm-corrected pairwise contrasts with effect sizes, Pearson's, Spearman's and polychoric correlation matrices, eigenvalues with parallel-analysis thresholds, and the rotated loading matrices, and the clustered re-analysis: intraclass correlations, the omnibus test on the nine group-mean profiles, and all 55 pairwise contrasts under the mixed model, under cluster-robust standard errors and under the cluster sign-flip permutation test; S3: Theme definitions and coding trajectory; and S4: COREQ checklist.

Author Contributions: Conceptualization, H.N.-W.; methodology, H.N.-W., L.D., I.S.V. and E.M.B.S.; formal analysis, H.N.-W.; investigation, H.N.-W., L.D., I.S.V. and E.M.B.S.; resources, H.N.-W., L.D.,

I.S.V. and E.M.B.S.; writing—original draft preparation, H.N.-W., L.D. and I.S.V.; writing—review and editing, H.N.-W., L.D., I.S.V. and E.M.B.S.; supervision, L.D. and I.S.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by BGDV-Hanbo Research under grant number BRI.16.R.161.RW.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Blithe Research Initiative (BRI), approval number BRI.581506.18215/1/1, issued 18 May 2023.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw rating matrix (45 participants × 11 criteria), the complete analysis script that reproduces every statistic and figure reported here, and all derived statistical output are provided in the Supplementary Materials, S2, published with this article. The workshop transcripts and field notes are retained but are not published as a corpus, and the reason is specific rather than general. The transcripts are a continuous record of design sessions in which participants proposed platform concepts, features and interface solutions that form the basis of ongoing commercial development by the authors; that material is interleaved with the material reported here rather than separable from it, so publishing the corpus would publish the design concepts with it. The restriction does not extend to individual participant quotations on topics that disclose no design concept, and this version accordingly reports additional verbatim extracts, principally on gamification, which the earlier version did not. Further quotation of that kind can be supplied to the editor or to reviewers on request. The participant-generated design artefacts and the paper prototypes were not retained at all, as described in Sections 3.6 and 5.4. Requests for restricted access for verification purposes may be directed to the corresponding author and will be considered on a case-by-case basis under a confidentiality agreement. An earlier account of this study is available as a preprint [23].

Acknowledgments: The authors acknowledge the partial use of generative AI and large language model tools for copy-editing, to improve language and readability, in accordance with emerging best practices in academic publishing. Such tools were not used for study design, data analysis or the interpretation of results. The manuscript was reviewed in full by the authors, who are solely responsible for its content, accuracy and integrity.

Conflicts of Interest: The authors declare the following conflicts of interest. First, the study's funder, BGDV-Hanbo Research, is affiliated with the Blithe Research Initiative, the body whose ethics committee granted approval for this study; the funder had no role in the design of the study, in the collection, analysis or interpretation of data, in the writing of the manuscript, or in the decision to publish. Second, the design concepts and materials generated in the co-design workshops reported here form the basis of ongoing commercial development work by the authors; the authors therefore have a commercial interest in the domain addressed by this study. Neither interest influenced the analysis or the reporting of results, and the quantitative analysis is provided in full and in reproducible form so that readers may verify this independently.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AMGR-I	Name of the prototype platform developed in this study (pilot I); not an initialism
BGDV	BGDV-Hanbo Research
BRI	Blithe Research Initiative
BWS	Best–Worst Scaling
CI	Confidence Interval
COREQ	Consolidated Criteria for Reporting Qualitative Research
DCE	Discrete Choice Experiment
DMHI	Digital Mental Health Intervention
HCI	Human–Computer Interaction

IQR	Interquartile Range
KMO	Kaiser–Meyer–Olkin measure of sampling adequacy
LLM	Large Language Model
MSA	Measure of Sampling Adequacy
N	Total sample size
NASSS	Non-adoption, Abandonment, Scale-up, Spread, and Sustainability framework
PCA	Principal Component Analysis
SD	Standard Deviation

References

1. Lipschitz, J.M.; Van Boxtel, R. ; Torous, J.; Firth, J.; Lebovitz, J.G.; Burdick, K.E.; Hogan, T.P. Digital Mental Health Interventions for Depression: Scoping Review of User Engagement. *J. Med. Internet Res.* **2022**, *24*, e39204. [CrossRef] [PubMed]
2. Fleming, T.; Bavin, L.; Lucassen, M.; Stasiak, K.; Hopkins, S.; Merry, S. Beyond the Trial: Systematic Review of Real-World Uptake and Engagement with Digital Self-Help Interventions for Depression, Low Mood, or Anxiety. *J. Med. Internet Res.* **2018**, *20*, e199. [CrossRef] [PubMed]
3. Valentine, L.; Hinton, J.D.X.; Bajaj, K.; Boyd, L.; O'Sullivan, S.; Sorenson, R.P.; Bell, I.H.; Vega, M.S.; Liu, P.; Peters, W.; et al. A meta-analysis of persuasive design, engagement, and efficacy in 92 RCTs of mental health apps. *npj Digit. Med.* **2025**, *8*, 229. [CrossRef] [PubMed]
4. Bond, R.; Mulvenna, M.; Potts, C.; O'Neill, S.; Ennis, E.; Torous, J. Digital transformation of mental health services. *npj Ment. Health Res.* **2023**, *2*, 13. [CrossRef] [PubMed]
5. Bucci, S.; Schwannauer, M.; Berry, N. The digital revolution and its impact on mental health care. *Psychol. Psychother. Theory Res. Pract.* **2019**, *92*, 277–297. [CrossRef] [PubMed]
6. Lattie, E.G.; Lipson, S.K.; Eisenberg, D. Technology and College Student Mental Health: Challenges and Opportunities. *Front. Psychiatry* **2019**, *10*, 246. [CrossRef] [PubMed]
7. Lattie, E.G.; Stiles-Shields, C.; Graham, A.K. An Overview of and Recommendations for More Accessible Digital Mental Health Services. *Nat. Rev. Psychol.* **2022**, *1*, 87–100. [CrossRef] [PubMed]
8. Kealy, C.; Potts, C.; Madrid-Cagigal, A.; Mulvenna, M.D.; Donohoe, G.; O'Neill, S.; Barry, M.M. Co-production of accessible digital mental health supports with marginalised young people: A scoping review. *J. Ment. Health* **2026**, *35*, 76–93. [CrossRef] [PubMed]
9. Balcombe, L.; De Leo, D. Digital Mental Health Challenges and the Horizon Ahead for Solutions. *JMIR Ment. Health* **2021**, *8*, e26811. [CrossRef] [PubMed]
10. Löchner, J.; Carlbring, P.; Schuller, B.W.; Torous, J.; Sander, L. Digital interventions in mental health: An overview and future perspectives. *Internet Interv.* **2025**, *40*, 100824. [CrossRef] [PubMed]
11. Sanders, E.; Stappers, P.J. Co-creation and the new landscapes of design. *Codesign* **2008**, *4*, 5–18. [CrossRef]
12. Bird, M.; McGillion, M.; Chambers, E.; Dix, J.; Fajardo, C.; Gilmour, M.; Levesque, K.; Lim, A.; Mierdel, S.; Ouellette, C.; et al. A generative co-design framework for healthcare innovation: Development and application of an end-user engagement framework. *Res. Involv. Engagem.* **2021**, *7*, 12. [CrossRef] [PubMed]
13. Duffy, A.; Boroumandzad, N.; Sherman, A.; Christie, G.; Riadi, I.; Moreno, S. Examining Challenges to Co-Design Digital Health Interventions with End Users: Systematic Review. *J. Med. Internet Res.* **2025**, *27*, e50178. [CrossRef] [PubMed]
14. Orlowski, S.; Matthews, B.; Bidargaddi, N.; Jones, G.; Lawn, S.; Venning, A.; Collin, P. Mental Health Technologies: Designing with Consumers. *JMIR Hum. Factors* **2016**, *3*, e4. [CrossRef] [PubMed]
15. Veldmeijer, L.; Terlouw, G.; Van Os, J.; Van Dijk, O.; Van 't Veer, J.; Boonstra, N. The Involvement of Service Users and People with Lived Experience in Mental Health Care Innovation Through Design: Systematic Review. *JMIR Ment. Health* **2023**, *10*, e46590. [CrossRef] [PubMed]
16. Moore, G.; Campbell, M.; Copeland, L.; Craig, P.; Movsisyan, A.; Hoddinott, P.; Littlecott, H.; O'Cathain, A.; Pfadenhauer, L.M.; Rehfuess, E.; et al. Adapting interventions to new contexts—The ADAPT guidance. *BMJ* **2021**, *374*, n1679. [CrossRef] [PubMed]
17. Ball, H.; Eisner, E.; Nicholas, J.; Wilson, P.; Bucci, S. How theories, models, and frameworks have been used to implement digital health interventions in services for people with severe mental health problems: A scoping review. *BMC Public Health* **2025**, *25*, 1023. [CrossRef] [PubMed]
18. Naderbagi, A.; Loblay, V.; Zahed, I.U.M.; Ekambareshwar, M.; Poulsen, A.; Song, Y.J.C.; Ospina-Pinillos, L.; Krausz, M.; Kamel, M.M.; Hickie, I.B.; et al. Cultural and Contextual Adaptation of Digital Health Interventions: Narrative Review. *J. Med. Internet Res.* **2024**, *26*, e55130. [CrossRef] [PubMed]
19. Slattery, P.; Saeri, A.K.; Bragge, P. Research co-design in health: A rapid overview of reviews. *Health Res. Policy Syst.* **2020**, *18*, 17. [CrossRef] [PubMed]

20. Vargas, C.; Longworth, G.R.; Zorbas, C.; Needham, C.; Ugalde, A.; Allender, S. Establishing fundamental co-design principles and strategies for public health research practice. *Codesign* **2026**, 1–22 . [CrossRef]
21. Steen, M. Co-Design as a Process of Joint Inquiry and Imagination. *Des. Issues* **2013**, *29*, 16–28. [CrossRef]
22. Cosgrave, C.; Kennedy, A.; Dietrich, T.; Gunn, K.M.; MacDonald, J.; McKay, C. ; Cunningham, S.; Haren, S.; Lewis, J. From co-design to co-production: Approaches, enablers, and constraints in developing a public health, capacity-building solution. *Aust. J. Rural Health* **2022**, *30*, 738–746. [CrossRef] [PubMed]
23. Ngabo-Woods, H. Insights Beyond Functionality: Empowering Digital Mental Health User Experiences Through Participatory Engagement and Evaluation. *Preprints* **2023**. [CrossRef]
24. Bødker, S.; Dindler, C.; Iversen, O.S.; Smith, R.C. What Is Participatory Design? In *Participatory Design* ; Synthesis Lectures on Human-Centered Informatics; Springer: Cham, Switzerland , 2022; pp. 5–13. [CrossRef]
25. Effendy, C.; Margaret, S.E.P.M.; Probandari, A. The Utility of Participatory Action Research in the Nursing Field: A Scoping Review. *Creat. Nurs.* **2022**, *28*, 54–60. [CrossRef] [PubMed]
26. Smith, R.C.; Bossen, C.; Kanstrup, A.M. Participatory design in an era of participation. *Codesign* **2017**, *13*, 65–69 . [CrossRef]
27. Van der Velden, M. ; Mörtberg, C. Participatory Design and Design for Values. In *Handbook of Ethics, Values, and Technological Design*; Springer: Dordrecht, The Netherlands, 2014; pp. 1–22. [CrossRef]
28. Sadek, M.; Calvo, R.A.; Mougenot, C. Co-designing conversational agents: A comprehensive review and recommendations for best practices. *Des. Stud.* **2023**, *89*, 101230. [CrossRef]
29. Naslund, J.A.; Bondre, A.; Torous, J.; Aschbrenner, K.A. Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. *J. Technol. Behav. Sci.* **2020**, *5*, 245–257. [CrossRef] [PubMed]
30. Gray, L.; Wong-Wylie, G.; Rempel, G.R.; Cook, K. Expanding Qualitative Research Interviewing Strategies: Zoom Video Communications. *Qual. Rep.* **2020**, *25*, 1292–1301. [CrossRef]
31. Porche, M.V.; Folk, J.B.; Tolou-Shams, M.; Fortuna, L.R. Researchers' Perspectives on Digital Mental Health Intervention Co-Design with Marginalized Community Stakeholder Youth and Families. *Front. Psychiatry* **2022**, *13*, 867460. [CrossRef] [PubMed]
32. Yardley, L.; Morrison, L.; Bradbury, K.; Muller, I. The Person-Based Approach to Intervention Development: Application to Digital Health-Related Behavior Change Interventions. *J. Med. Internet Res.* **2015**, *17*, e30. [CrossRef] [PubMed]
33. Skivington, K.; Matthews, L.; Simpson, S.A.; Craig, P.; Baird, J.; Blazeby, J.M.; Boyd, K.A.; Craig, N.; French, D.P.; McIntosh, E.; et al. A new framework for developing and evaluating complex interventions: Update of Medical Research Council guidance. *BMJ* **2021**, *374*, n2061. [CrossRef] [PubMed]
34. Greenhalgh, T.; Wherton, J.; Papoutsi, C.; Lynch, J.; Hughes, G.; A'Court, C.; Hinder, S.; Fahy, N.; Procter, R.; Shaw, S. Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. *J. Med. Internet Res.* **2017**, *19*, e367. [CrossRef] [PubMed]
35. Mohr, D.C.; Lyon, A.R.; Lattie, E.G.; Reddy, M.; Schueller, S.M. Accelerating Digital Mental Health Research From Early Design and Creation to Successful Implementation and Sustainment. *J. Med. Internet Res.* **2017**, *19*, e153. [CrossRef] [PubMed]
36. Ryan, M.; Gerard, K.; Amaya-Amaya, M. *Using Discrete Choice Experiments to Value Health and Health Care*; Springer: Dordrecht, The Netherlands, 2008. [CrossRef]
37. Cheung, K.L.; Wijnen, B.F.M.; Hollin, I.L.; Janssen, E.M.; Bridges, J.F.; Evers, S.M.A.A.; Hiligsmann, M. Using Best–Worst Scaling to Investigate Preferences in Health Care. *PharmacoEconomics* **2016**, *34*, 1195–1209. [CrossRef] [PubMed]
38. Ferri Sanz, M.; Vallina Acha, B.; Ferrando García, M.F. Co-Design for People-Centred Care Digital Solutions: A Literature Review. *Int. J. Integr. Care* **2021**, *21*, 16. [CrossRef] [PubMed]
39. Silverio, S.A.; Sheen, K.; Bramante, A.; Knighting, K.; Koops, T.; Montgomery, E.; November, L.; Soulsby, L.K.; Stevenson, J.H.; Watkins, M.; et al. Sensitive, Challenging, and Difficult Topics: Experiences and Practical Considerations for Qualitative Researchers. *Int. J. Qual. Methods* **2022**, *21*, 1–16. [CrossRef]
40. McCombie, C.; Miguel Esponda, G. ; Ouazzane, H.; Knowles, G.; Gayer-Anderson, C.; Schmidt, U.; Lawrence, V. Qualitative Digital Diary Methods: Participant-Led Values for Ethical and Insightful Mental Health Research. *Int. J. Qual. Methods* **2024**, *23*, 1–11. [CrossRef] [PubMed]
41. Hassenzahl, M.; Diefenbach, S.; Göritz, A.S. Needs, affect, and interactive products—Facets of user experience. *Interact. Comput.* **2010**, *22*, 353–362. [CrossRef]
42. Bødker, S. When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles*; ACM: New York, NY, USA, 2006; pp. 1–8. [CrossRef]
43. Fortuna, K.L.; Hill, J.; Chalker, S.A.; Ferron, J.C. Certified Peer Support Specialists Training in Technology and Delivery of Digital Peer Support Services: Cross-Sectional Study. *JMIR Form. Res.* **2022**, *6*, e40065. [CrossRef] [PubMed]
44. Fortuna, K.L.; Naslund, J.A.; LaCroix, J.M.; Bianco, C.; Brooks, J.; Zisman-Ilani, Y.; Muralidharan, A.; Deegan, P.E. Digital Peer Support Mental Health Interventions for People with a Lived Experience of a Serious Mental Illness: Systematic Review. *JMIR Ment. Health* **2020**, *7*, e16460. [CrossRef] [PubMed]

45. Coelho, J.; Pécune, F.; Micoulaud-Franchi, J.; Bioulac, B.; Philip, P. Promoting mental health in the age of new digital tools: Balancing challenges and opportunities of social media, chatbots, and wearables. *Front. Digit. Health* **2025**, *7*, 1560580. [CrossRef] [PubMed]
46. Torous, J.; Linardon, J.; Goldberg, S.B.; Sun, S.; Bell, I.; Nicholas, J.; Hassan, L.; Hua, Y.; Milton, A.; Firth, J. The evolving field of digital mental health: Current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry* **2025**, *24*, 156–174. [CrossRef] [PubMed]
47. Lipschitz, J.M.; Pike, C.K.; Hogan, T.P.; Murphy, S.A.; Burdick, K.E. The Engagement Problem: A Review of Engagement with Digital Mental Health Interventions and Recommendations for a Path Forward. *Curr. Treat. Options Psychiatry* **2023**, *10*, 119–135. [CrossRef] [PubMed]
48. Valentine, L.; Nicholas, J.; Sorenson, R.P.; Chen, N.A.; McEnery, C.; Louis, S.; Cross, S.; Mangelsdorf, S.N.; O'Sullivan, S.; Wren, T.W.; et al. Sustained engagement with a digital youth mental health platform: A mixed-methods study. *Internet Interv.* **2026**, *43*, 100899. [CrossRef] [PubMed]
49. Schiffler, T.; Kapan, A.; Gansterer, A.; Paß, T.; Lehner, L.; Gil-Salmerón, A.; McDermott, D.T.; Grabovac, I. Characteristics and Effectiveness of Co-Designed Mental Health Interventions in Primary Care for People Experiencing Homelessness: A Systematic Review. *Int. J. Environ. Res. Public Health* **2023**, *20*, 892. [CrossRef] [PubMed]
50. Creswell, J.W.; Plano Clark, V.L. *Designing and Conducting Mixed Methods Research*, 3rd ed.; SAGE Publications: Thousand Oaks, CA, USA, 2018.
51. Fetters, M.D.; Curry, L.; Creswell, J.W. Achieving Integration in Mixed Methods Designs—Principles and Practices. *Health Serv. Res.* **2013**, *48*, 2134–2156. [CrossRef] [PubMed]
52. Kaiser, H.F. An index of factorial simplicity. *Psychometrika* **1974**, *39*, 31–36. [CrossRef]
53. Horn, J.L. A rationale and test for the number of factors in factor analysis. *Psychometrika* **1965**, *30*, 179–185. [CrossRef] [PubMed]
54. Braun, V.; Clarke, V. Reflecting on reflexive thematic analysis. *Qual. Res. Sport Exerc. Health* **2019**, *11*, 589–597. [CrossRef]
55. Byrne, D. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Qual. Quant.* **2022** , *56*, 1391–1412. [CrossRef]
56. Tong, A.; Sainsbury, P.; Craig, J. Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *Int. J. Qual. Health Care* **2007**, *19*, 349–357. [CrossRef] [PubMed]
57. Lincoln, Y.S.; Guba, E.G. *Naturalistic Inquiry*; Sage Publications: Beverly Hills, CA, USA, 1985.
58. Nowell, L.; Norris, J.M.; White, D.; Moules, N.J. Thematic Analysis: Striving to Meet the Trustworthiness Criteria . *Int. J. Qual. Methods* **2017**, *16*, 1–13. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.