




# A process for reviewing mental health apps: Using the One Mind PsyberGuide Credibility Rating System

Digital Health  
Volume 7: 1–10  
© The Author(s) 2021  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20552076211053690  
journals.sagepub.com/home/dhj  


Martha Neary<sup>1</sup> , John Bunyi<sup>1</sup>, Kristina Palomares<sup>1</sup>, David C. Mohr<sup>2</sup>, Adam Powell<sup>3</sup> , Josef Ruzek<sup>4,5</sup>, Leanne M. Williams<sup>5</sup>, Til Wykes<sup>6,7</sup> and Stephen M. Schueller<sup>1</sup>

## Abstract

**Objective:** Given the increasing number of publicly available mental health apps, we need independent advice to guide adoption. This paper discusses the challenges and opportunities of current mental health app rating systems and describes the refinement process of one prominent system, the One Mind PsyberGuide Credibility Rating Scale (PGCRS).

**Methods:** PGCRS Version 1 was developed in 2013 and deployed for 7 years, during which time a number of limitations were identified. Version 2 was created through multiple stages, including a review of evaluation guidelines and consumer research, input from scientific experts, testing, and evaluation of face validity. We then re-reviewed 161 mental health apps using the updated rating scale, investigated the reliability and discrepancy of initial scores, and updated ratings on the One Mind PsyberGuide public app guide.

**Results:** Reliabilities across the scale's 9 items ranged from  $-0.10$  to  $1.00$ , demonstrating that some characteristics of apps are more difficult to rate consistently. The average overall score of the 161 reviewed mental health apps was  $2.51/5.00$  (range  $0.33-5.00$ ). Ratings were not strongly correlated with app store star ratings, suggesting that credibility scores provide different information to what is contained in star ratings.

**Conclusion:** PGCRS summarizes and weights available information in 4 domains: intervention specificity, consumer ratings, research, and development. Final scores are created through an iterative process of initial rating and consensus review. The process of updating this rating scale and integrating it into a procedure for evaluating apps demonstrates one method for determining app quality.

## Keywords

mHealth, mental health, mobile health, evaluation, digital mental health

Submission date: 9 April 2021; Acceptance date: 29 September 2021

## Introduction

### *Mental health apps: opportunities and challenges*

The increasing availability of technologies, such as smartphones, affords opportunities to increase access to mental health care. These technologies are more crucial than ever in the era of COVID-19, when mental health concerns are increased and additional, unique barriers to care exist (such as physical distancing measures which limit contact with providers).<sup>1</sup> An estimated 325,000 mobile health apps are available in the app marketplace,<sup>2</sup> with at least

<sup>1</sup>Department of Psychological Science, University of California, University of California, Irvine, CA, USA

<sup>2</sup>Center for Behavioral Intervention Technologies, Department of Preventive Medicine, Northwestern University, Chicago, IL, USA

<sup>3</sup>Payer+Provider Syndicate, Boston, MA, USA

<sup>4</sup>Palo Alto University, Palo Alto, CA, USA

<sup>5</sup>Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA

<sup>6</sup>Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

<sup>7</sup>South London and Maudsley NHS Foundation Trust London, UK

#### Corresponding author:

Martha Neary, University of California, Irvine, Department of Psychological Science, 4201 Social and Behavioral Sciences Gateway, Irvine, CA, USA.  
Email: mneary1@uci.edu



10,000 of those for mental health.<sup>3</sup> Consumer interest in mental health apps (MH apps) is growing; 64% of teens and young adults report using health apps, with many of those apps being related to mental health including sleep, meditation, stress, and substance use.<sup>4</sup>

Despite reported interest, those wanting to use MH apps often have little help in selecting potentially effective products. The app stores only provide star ratings, and these user reviews correlate poorly with clinical utility.<sup>5</sup> People who contribute app ratings are a self-selected sample likely to represent technologically-savvy users<sup>5</sup> or those with a particularly negative or positive experience to share.<sup>6</sup> App developers can also leave ratings for their own apps or pay others to do so and there is no way to distinguish genuine consumer ratings from those that are fraudulent.<sup>7</sup>

While thousands of MH apps are available, it is well documented that few have been reviewed, researched, or vetted in any systematic way.<sup>6,8-10</sup> In 2017, Firth and colleagues<sup>11</sup> conducted a systematic search of seven databases and identified only 18 randomized controlled trials (RCTs) examining the effects of mental health interventions for depression delivered via smartphones. A similar study in the same year identified only nine RCTs for anxiety apps.<sup>12</sup> While RCTs are the gold standard and would provide useful information for making choices, they are not available for every app, and the incentives for developers to complete trials are misaligned with incentives for researchers.<sup>13</sup> Consumers now increasingly look to professionals and “trusted sources” for app recommendations<sup>14</sup> which means that we need frameworks for rigorous evaluations.<sup>15</sup>

### *Navigating the MH app marketplace: some potential solutions*

In navigating the MH app marketplace, two common questions exist: “which apps are effective?”, and “how does one distinguish a good app from a bad app?”. Efforts to help people answer these questions can be broadly categorized into evaluation guidelines and app rating platforms.<sup>6,16</sup> Evaluation guidelines, for example the Mobile App Rating Scale (MARS),<sup>17</sup> the American Psychiatric Association’s App Evaluation Model,<sup>3</sup> and Enlight,<sup>18</sup> aim to guide the consumer through a number of questions to decide whether or not to proceed with using an app. However, these frameworks do not provide clear metrics to guide app choices. Even with the help of evaluation guidelines, consumers (even clinician consumers) generally do not have the time or qualifications to thoroughly evaluate apps.<sup>19</sup> This is despite recent efforts to make these guidelines more streamlined or provide additional materials to support their use.<sup>20</sup> These sorts of guidelines require careful consideration and evaluation of apps for security, credibility, and clinical efficacy, and so will be even more challenging for lay consumers, who generally want

simpler information to make choices.<sup>9,21</sup> Third-party quality reviews might fill this gap by providing information on the quality of an app at the point of download (e.g. on the app stores).<sup>21</sup> In the absence of such a solution, independent app rating platforms for smartphone apps that produce scores can help consumers and clinicians distinguish high-quality apps. These include the Organization for the Review of Care & Health Applications (ORCHA), MindTools.io, Credible Mind, and One Mind PsyberGuide, but these too have drawbacks.

### *Improving existing solutions*

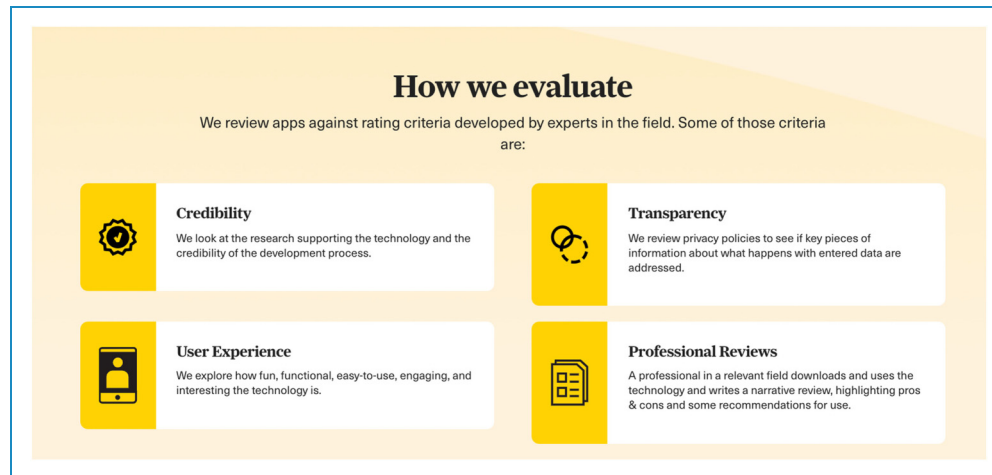
Recent work by Carlo and colleagues<sup>22</sup> demonstrated inconsistencies across different rating systems. They found low rating agreement for the most commonly downloaded MH and wellness apps by ORCHA, MindTools, and One Mind PsyberGuide. Ratings of credibility and evidence base demonstrated the most agreement, with ratings of user experience the least. Powell and colleagues<sup>19</sup> also found poor inter-rater reliability using the same measures, particularly for ratings of effectiveness. This “inherent methodological subjectivity” must be acknowledged,<sup>22</sup> and rating developments need to define criteria clearly to ensure consistency.<sup>19</sup>

### *The One Mind PsyberGuide Credibility Rating Scale*

One Mind PsyberGuide (hereafter “PsyberGuide”) is a non-profit organization providing reviews of digital tools (including both apps and web-based programs) for mental health and wellness. All reviews are publicly available at <https://onemindpsyberguide.org/>. In addition to narrative reviews by professionals, PsyberGuide reviews digital tools on three different metrics (shown in Figure 1) which map onto key considerations for service users<sup>23,24</sup> to help them make informed decisions. Although all three metrics might affect user adoption and engagement, in this paper we focus only on the PsyberGuide Credibility Rating Scale (PGCRS). The other measures have been described and evaluated elsewhere.<sup>17,25</sup> This paper describes the process of updating the PGCRS to better reflect the evidence and support backing MH apps.

The PGCRS is completed by a trained app reviewer for each tool. This rating is reviewed and discussed with a supervision team comprising two Master’s-level staff members and one PhD-level clinical psychologist. Final scores are based on discussion with the supervision team, and the maximum number of points possible for any tool is five.

The first version of the PGCRS (PCGRS 1.0) was created in 2013 and used for seven years, with some minor periodic updates. Informal feedback from consumers, developers, and researchers on key aspects of the original scale demonstrated that it did not capture all the



**Figure 1.** One Mind PsyberGuide evaluation metrics, as listed on onemindpsyberguide.org.

information that would be useful. Version 2 of the PGCRS (PGCRS 2.0) was developed to respond to these issues.

## Methods

PGCRS 2.0 development followed a series of stages, outlined in Figure 2 and explained in more detail below.

### Stage 1: Discovery

We reviewed available app rating frameworks, for example Enlight,<sup>18</sup> the American Psychiatric Association App Evaluation Model,<sup>3</sup> and ORCHA.<sup>26</sup> We also reviewed consumer research to understand what additional consumer questions pertaining to issues of credibility were not addressed by PGCRS 1.0.<sup>14,21,27,28</sup> Finally, we reviewed anecdotal feedback we have received on PGCRS 1.0 over the course of its implementation. The preliminary PGCRS 2.0 was then developed based on this evidence.

### Stage 2: Initial testing of PGCRS 2.0

Three experienced raters, who had completed dozens of ratings using PGCRS 1.0, used PGCRS 2.0 to review 10 apps. This process produced further clarifications to the wording and criteria (for example adding examples to the anchors for clarity of purpose; see Appendix for full rating tool).

### Stage 3: Expert input

PGCRS 2.0 was reviewed by the PsyberGuide Scientific Advisory Board, including all co-authors of this paper, to assess face validity. Based on their feedback we added items on indirect research evidence, development processes, efficacy of other products by the same development team, and the average value of consumer ratings. Details of

the item changes from PGCRS 1.0 to PGCRS 2.0 based on these three stages are presented in the results.

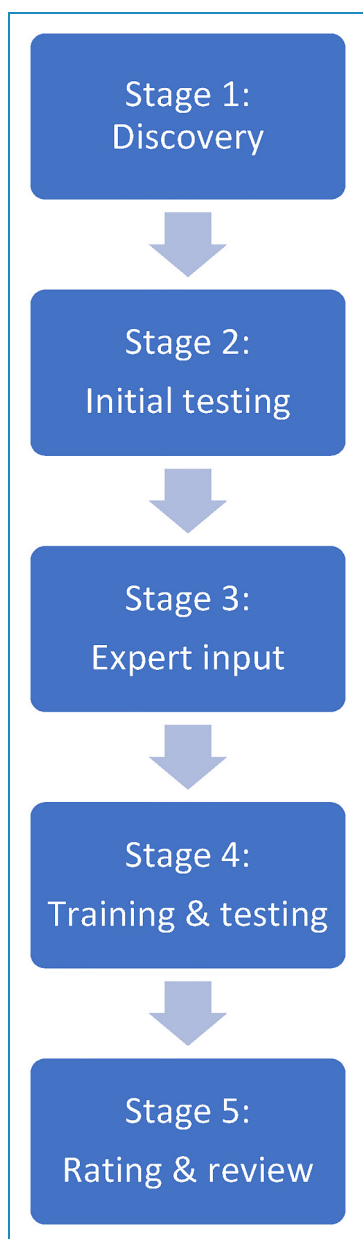
### Stage 4: Training and testing

Reviewers were three undergraduates and two graduates. They completed six weeks of training by an experienced team in digital mental health (including two graduate-level trained app reviewers and one clinical psychologist). During training they rated 15 training apps which involved downloading the assigned app, using it for a period of at least two hours across more than one day, and then producing an initial rating. Training apps were completed in batches of five. After each batch, reviewers and their supervisors met to review the initial scores, answer questions, and discuss the experience. In these meetings a consensus (final) score was determined for each of the 15 apps by resolving discrepancies through discussion.

To understand scoring differences between raters and the reliability of initial scores, discrepancies between the raters' initial scores and the app's consensus scores were examined. For each of the 15 training apps, consensus scores were subtracted from initial individual rater scores for each subscale. Inter-rater reliability of the initial scores was determined by calculating Krippendorff's Alpha using the 'R' statistical computing tool script provided by Zapf and colleagues.<sup>29</sup> Inter-rater reliability of the final scores could not be calculated because these scores were produced through a consensus process.

### Stage 5: Rating and review

Once training was completed, reviewers used each of the remaining available apps from the PsyberGuide App Guide ( $N=146$ ) and completed the PGCRS 2.0 (one rating per tool). Final scores were completed using a



**Figure 2.** Process of rating scale development.

consensus process. Including training apps, this resulted in 161 rated platforms in total, completed over a period of five and a half months. We compared the consensus scores between PGCRS 1.0 and 2.0 and investigated them in detail if the ratings changed by one point or greater ( $N=36$ ) or if they were in the top 10% ( $N=17$ ) or the bottom 10% ( $N=16$ ) of all ratings. This investigation was carried out by an experienced supervisor, who downloaded and used the tool, and examined the ratings, and approved the final score. If this supervisor had questions or was unsure of the score change, an additional supervisor also reviewed and discussed in order to reach a decision.

### *Correlating app store scores and credibility scores*

When all reviews were completed and approved, we examined the correlations between the app store star ratings and the PGCRS 2.0 scores, for tools that had a smartphone app available in either the Apple App Store or Google Play Store ( $N=147$ ). The star ratings (range 1 to 5) were obtained from the iOS and Android app stores using AppTrace, an analysis service which programmatically queries both iTunes and Google Play application programming interfaces (APIs). Mirroring the method used by Singh et al.<sup>5</sup> we queried the cumulative star rating from all previous versions of the app, instead of the summary rating for the current version only which is presented in the app store. As noted by Singh et al.<sup>5</sup> the rating from all versions represents a more stable estimate of an app's perceived value. For multiplatform apps, we calculated a mean rating based on the iOS and Android star ratings. Because the PGCRS 2.0 accounts for consumer ratings, we ran two correlations: (1) app store star rating and total PGCRS 2.0 score, and (2) app store star rating and PGCRS 2.0 score, minus the consumer rating.

## **Results**

### *Stages 1–3: Discovery, initial testing, and expert input*

The main features assessed by the Scale, and changes from PGCRS 1.0 to PGCRS 2.0 (made in Stages 1–3 of development), are shown in Table 1. The full rating tool and scoring are provided in the Appendix.

### *Stage 4: Training and testing*

To understand scoring differences between raters and reliability of initial scores, discrepancies between the raters' initial scores and the app's consensus scores were examined. Average discrepancies, standard deviation (SD), mean absolute error, and Krippendorff's Alpha are presented in Table 2. Because only four reviewers completed the last batch of apps, Krippendorff's Alphas are presented as ranges between the first 10 and last five apps.

### *Stage 5: Rating and review*

Of the 177 tools listed on the PsyberGuide App Guide, 15 (9%) were identified as no longer available (e.g. had been removed from the app store) leaving 161 tools currently available to the public. For ratings using PGCRS 2.0, the average overall score for the 161 tools was 2.51 (range 0.33–5.00;  $SD=1.23$ ) and compared to PGCRS 1.0, 42% ( $n=67$ ) increased their score and 58% ( $n=93$ ) decreased. The average score change was small ( $-0.04$ ) although some did show large changes (e.g. 2.24). Score changes

**Table 1.** Features assessed by the PsyberGuide Credibility Rating Scale (PGCRS).

Domain	Feature	New in V2	Rationale for addition
(1) Intervention Specificity	a. Clarity of proposed goal	✓	Without clear, measurable, specific goals, it is difficult to evaluate the success of a tool in meeting those goals. Goals should not only be clear, but achievable; a tool which over promises or makes lofty claims is unlikely to deliver on those goals (for example, “become more successful” or change your life”).
(2) Consumer Ratings	a. (i) Number of app store ratings		
	a. (ii) Average value	✓	In addition to the number of consumer reviews, which serves as a proxy for popularity, the average value of reviews can help distinguish apps which consumers rate highly or poorly.
(3) Research	a. Direct research evidence		
	b. Indirect research evidence	✓	There is value in a tool being grounded in indirect evidence and evidence-based practices. Ideally, tools will have both direct and indirect evidence.
	c. Research independence & review		
(4) Development	a. Development processes (e.g. pilot, feasibility & acceptability data; stakeholder engagement)	✓	Valuable lessons can be learned through data collected during the initial development or piloting process which can inform product development. It is also important for developers to solicit feedback from the stakeholders on what the app is designed to help, to ensure the app is meaningful and relevant to them.
	b. Efficacy of other products by same development team	✓	There is some knowledge and learning that comes from developing a previous MH app which can assist in the development of a subsequent product. Importantly, in order to receive points here the previous app needs to have been shown to be effective. This is consistent with FDA’s pre-certification process that incorporates developmental process and team into its review.
	c. Clinical input in development		
	d. Ongoing maintenance and updates (date of last software update)		

were attributable most commonly due to the number and average score of consumer ratings and ongoing maintenance and updates (date of last software update), which regularly fluctuate.

$r_s(145) = .18, p = 0.024$ . We also examined the correlation between app store star ratings and PCGRS 2.0 scores minus the consumer rating item. There was no significant correlation between the two,  $r_s(145) = .08, p = 0.268$ .

### Correlating app store scores and credibility scores

Spearman’s rho correlation coefficient was used to examine the relationship between app store star ratings and total PCGRS 2.0 final scores, and showed a small correlation,

### Discussion

This paper reports on the development and face validity of PCGRS 2.0 and its application to all available digital tools for mental health and wellness listed on the PsyberGuide

**Table 2.** Discrepancies between final and initial scores and inter-rater reliability for initial scores for training apps.

Domain	1			2			3			4		
Feature	a	a	a	b	c	a	b	c	a	b	c	d
<b>Average Discrepancy</b>	-0.26	0.00	0.00	-0.14	-0.09	-0.17	0.07	-0.17	0.00			
<b>SD</b>	0.53	0.00	0.54	0.39	0.53	0.54	0.31	0.38	0.00			
<b>Mean Absolute Error</b>	0.31	0.00	0.23	0.17	0.20	0.31	0.10	0.17	0.00			
<b>Kripp. alpha</b>	0.42-0.62	1.00	0.63-0.77	-0.39-0.15	0.63-0.67	0.10-0.18	0.68-0.81	-0.10-0.19	1.00			

Note: Feature numbers correspond to those listed in Table 1.

App Guide at onemindpsyberguide.org. During the review process, nearly a tenth of digital tools listed on PsyberGuide were identified as no longer available (and were moved to the “currently unavailable” section of the guide). This speaks to the rapidly changing marketplace in which consumers and clinicians search for suitable apps, supporting our view that app recommendations are vital, but also demonstrating the challenge of making these ratings current enough to enable consumer choices.

The final scale was reasonably reliable and provided a measure to assess the credibility of MH apps when completed by trained raters (i.e. undergraduate students) under supervision. It includes updated and additional items deemed important through a review of available frameworks, relevant literature, feedback from developers, and input from subject matter experts. We also chose specifically to embed factors identified in the literature as important to consumers, such as development processes and feasibility data,<sup>21</sup> direct research evidence and evidence-based content, and clinical input in development.<sup>14,27,28</sup>

Our rating process included reviewers producing initial scores using the PGCRS and then creating final scores through a consensus process with discussion and supervision. Even after training, ratings were regularly discussed in team meetings to allow opportunities to calibrate scores, provide ongoing supervision, and produce final scores. Reviewing the discrepancies between the raters’ initial scores and the app’s final (consensus) scores (Table 2) allowed us to identify those items requiring further discussion.

Using 0.667 as an acceptable level of reliability,<sup>30</sup> five items had good reliability (items 2a, 3a, 3c, 4b, 4d) and four items had relatively low reliability (items 1a, 3b, 4a, 4c). Discrepancies for low reliability items were likely due to sparse or hard-to-find information; for example, 1) information on stakeholder or consumer involvement is not always readily available (item 4a), and 2) it is not always easy to track new products from the same development team, with frequent changes to company names, app names, and websites (item 4b). Items that required

discussion included whether a development team could be counted as clinical input (item 4c). High reliability items were those where objective information was available on app stores or publicly available databases, for example, the number of consumer ratings (item 2a), direct research evidence (item 3a), and date of last software update (item 4d). In future these low reliability items will be refined to include more concrete anchors or examples. As it is emphasized now in app development the amount of clinical input or consumer involvement should become more transparent.

It was unsurprising that ratings showed poor correlations with app store star ratings, replicating previous research that star ratings are not predictive of app quality,<sup>5,19</sup> and that additional ratings beyond app store information are needed to guide consumer and clinician choices.

Although many app evaluation models and rating systems have been developed over the last decade, few have incorporated the views or needs of patients and consumers. Consumers are ultimately the end users of all MH apps, and factors which influence consumer adoption and use of apps do not necessarily align with the views of expert or academic groups. Wykes and Schueller<sup>21</sup> propose that information consumers need to make app choices falls into four domains, which were derived from experimental studies, systematic reviews, and reports of patient concerns: (1) privacy and data security, (2) development characteristics, (3) feasibility data, and (4) benefits (the first concern is addressed by the PsyberGuide Transparency Rating, while concerns 2–4 are addressed by PGCRS 2.0 which is the focus of the current paper). More work is needed to ensure that consumer perspectives are central to MH app choices and that we integrate both “bottom up” (consumer-informed) and “top down” (expert-driven) processes in evaluation.

Our efforts to re-rate all tools on the PsyberGuide App Guide demonstrate the importance of regular updating. Changes in scores reflect not only the updated scale, but also changes in information that informs the scores, such as additional research. This demonstrates the need for app

ratings to be nimble and regularly assess whether new information will affect those scores. As more MH apps become available, the challenge of keeping reviews up to date will grow more arduous. We agree with calls for continuous, real-time evaluation of apps to guide evaluation efforts.<sup>13,31</sup> However, to date, there is no process through which third-party app evaluators can obtain real-world effectiveness data for multiple products, and in the absence of such an infrastructure, expert reviews which are regularly updated are likely the best current solution.

Echoing Powell and colleagues,<sup>9</sup> we believe it is problematic to ask clinicians and patients to fend for themselves when evaluating apps. For app ratings to be truly informative and useful, they need to come from objective, unbiased, third-party reviewers, independent from commercial app development efforts. The necessity for independent app evaluation systems has only been heightened by COVID-19, due to both the increased need for digital supports for mental health<sup>1</sup> and further loosening of FDA regulation in order to expand the availability of digital health therapeutic devices for patient and consumer use.<sup>32</sup>

### Strengths and limitations

The PGCRS 2.0 is a measure to determine app quality that is different to what is provided via the app stores and star ratings. However, it is worth noting some limitations of the current investigation and the scale. We have not carried out a measure validation study. No gold standard measure of app quality exists or is widely available. Therefore, we cannot validate the accuracy of the scale in predicting app quality. The clinical validity of the PGCRS would require examining whether it correlates with clinical benefits, but as of now, no repository of such information exists. Scores resulting from PGCRS 1.0 ratings have been shared in various studies and contexts,<sup>19,33</sup> and by various organizations including the Anxiety and Depression Association of America and the International Obsessive Compulsive Disorder Foundation, suggesting acceptable face validity of the ratings. We have also responded to consumer feedback in the development of PGCRS 2.0. However, the only item that directly considers consumer input is item 2a (number of consumer ratings in the app store and the average star rating). Further considerations should be given to how to incorporate more informative consumer-driven approaches.

### Conclusion

The PGCRS 2.0 presents one evaluative framework to quantify the quality of a MH app through the lens of credibility. This credibility metric includes considerations of research evidence (both direct and indirect), the developmental processes, intervention specificity, and consumer ratings. The credibility

metric is meant to simplify and weight information available on MH apps in a manner such can be used by consumers – both professionals and non-professionals – to guide decision-making. The PGCRS underlies one aspect of the rating system used at PsyberGuide, which has been an influential system in rating apps, demonstrated through its use by various organizations. The process of updating the PGCRS from Version 1.0 to 2.0 also illustrates some important considerations as the field of digital mental health has developed. This includes incorporating indirect evidence given the growing evidence-base in this field and developmental processes. Although no system is perfect, this description and analysis helps demonstrate some of the strengths and limitations of this metric including highlighting the usefulness of embedding metrics into a consensus process. Better transparency around different evaluative frameworks used in this space will hopefully help drive the field forward and improve access to information that can help all stakeholders make informed decisions.

**Acknowledgements:** We would like to thank Zoe Dodge-Rice, Phatthawit Ketsing, Isabelle Lee, Lisa Vasquez, and Zhengxin Wan who were part of the app review team.

**Declaration of conflicting interests:** SMS receives funding from One Mind for the operation and management of One Mind PsyberGuide. SMS has also received unrelated consulting payments from Otsuka Pharmaceuticals. DCM, AP, JR, LW, and TW are members of the One Mind PsyberGuide Scientific Advisory Board for which they receive compensation.

**Funding:** One Mind funds the operation and management of One Mind PsyberGuide. This paper received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. TW would like to acknowledge funding from the NIHR Maudsley BRC and her NIHR Senior Investigator Award.

**Guarantor:** MN

**Contributorship:** MN and SMS conceived the study. JB oversaw review process and data analysis, and drafted the method and results. KP researched literature and drafted the introduction. MN wrote the first draft of the manuscript, with support and revisions from SMS. SMS, DCM, AP, JR, LW, and TW all provide feedback on paper drafts. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

**ORCID iDs:** Martha Neary  <https://orcid.org/0000-0002-1253-3459>

Adam Powell  <https://orcid.org/0000-0001-6519-3120>

### References

1. Liu S, Yang L, Zhang C, et al. Online mental health services in China during the COVID-19 outbreak. *Lancet Psychiat* 2020; 7: e17–e18.

2. Research2Guidance. 325,000 mobile health apps available in 2017 – Android now the leading mHealth platform, <https://research2guidance.com/325000-mobile-health-apps-available-in-2017/> (2017, accessed 18 December 2020).
3. Torous JB, Chan SR, Gipson SY-MT, et al. A hierarchical framework for evaluation and informed decision making regarding smartphone apps for clinical care. *Psychiatr Serv* 2018; 69: 498–500.
4. Rideout, Victoria; Fox, Susannah; and Well Being Trust. Digital Health Practices, Social Media Use, and Mental Well-Being Among Teens and Young Adults in the U.S. Report, *Articles, Abstracts, and Reports*, 1093, Summer 2018. <https://digitalcommons.psjhealth.org/publications/1093>
5. Singh K, Drouin K, Newmark LP, et al. Many Mobile health apps target high-need, high-cost populations, But gaps remain. *Health Aff (Millwood)* 2016; 35: 2310–2318.
6. Neary M and Schueller SM. State of the field of mental health apps. *Cogn Behav Pract*. 2018; 25(4), 531–537. <https://doi.org/10.1016/j.cbpra.2018.01.002>
7. BinDhim NF, Hawkey A and Trevena L. A systematic review of quality assessment methods for smartphone health apps. *Telemed E-Health* 2015; 21: 97–104.
8. Leigh S and Flatt S. App-based psychological interventions: friend or foe? *Evid Based Ment Health* 2015; 18: 97–99.
9. Powell AC, Landman AB and Bates DW. In search of a Few good apps. *JAMA* 2014; 311: 1851–1852.
10. Torous J and Powell AC. Current research and trends in the use of smartphone applications for mood disorders. *Internet Interv* 2015; 2: 169–173.
11. Firth J, Torous J, Nicholas J, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. *World Psychiatry Off J World Psychiat Assoc WPA* 2017; 16: 287–298.
12. Firth J, Torous J, Nicholas J, et al. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. *J Affect Disord* 2017; 218: 15–22.
13. Mohr DC, Cheung K, Schueller SM, et al. Continuous evaluation of evolving behavioral intervention technologies. *Am J Prev Med* 2013; 45: 517–523.
14. Schueller SM, Neary M, O’Loughlin K, et al. Discovery of and interest in health apps Among those With mental health needs: survey and focus group study. *J Med Internet Res* 2018; 20: e10141.
15. NIMH Opportunities and Challenges of Developing Information Technologies on Behavioral and Social Science Clinical Research. <https://www.nimh.nih.gov/about/advisory-boards-and-groups/namhc/namhc-workgroups/namhc-bssr-workgroup-charge.shtml> (accessed 3 January 2021).
16. Torous J, Powell A and Knable MB. Quality assessment of self-directed software and Mobile applications for the treatment of mental illness. *Psychiatr Ann* 2016; 46: 579–583.
17. Stoyanov SR, Hides L, Kavanagh DJ, et al. Mobile App Rating Scale: a new tool for assessing the quality of health mobile apps. *JMIR MHealth UHealth* 2015; 3(1):e27, <https://mhealth.jmir.org/2015/1/e27/>
18. Baumel A, Faber K, Mathur N, et al. Enlight: a comprehensive quality and therapeutic potential evaluation tool for Mobile and Web-based eHealth interventions. *J Med Internet Res* 2017; 19(3):e82, <https://www.jmir.org/2017/3/e82/>
19. Powell AC, Torous J, Chan S, et al. Interrater reliability of mHealth App rating measures: analysis of Top depression and smoking cessation apps. *JMIR MHealth UHealth* 2016; 4: e15.
20. APA App Advisor. <https://www.psychiatry.org/psychiatrists/practice/mental-health-apps> (accessed 18 December 2020).
21. Wykes T and Schueller S. Why reviewing apps Is Not enough: transparency for trust (T4 T) principles of responsible health App marketplaces. *J Med Internet Res* 2019; 21: e12390.
22. Carlo AD, Ghomi RH, Renn BN, et al. By the numbers: ratings and utilization of behavioral health mobile applications. *Npj Digit Med* 2019; 2(1):54, <https://doi.org/10.1038/s41746-019-0129-6>
23. Lipschitz J, Miller CJ, Hogan TP, et al. Adoption of mobile apps for depression and anxiety: Cross-sectional survey study on patient interest and barriers to engagement. *JMIR Ment Health* 2019; 6(1):e11334, <https://doi.org/10.2196/11334>
24. Freeman D, Sheaves B, Goodwin GM, et al. The effects of improving sleep on mental health (OASIS): a randomised controlled trial with mediation analysis. *Lancet Psychiatry* 2017; 4: 749–758.
25. O’Loughlin K, Neary M, Adkins EC, et al. Reviewing the data security and privacy policies of mobile apps for depression. *Internet Interv*. 2018; 15: 110-115, <https://doi.org/10.1016/j.invent.2018.12.001>
26. ORCHA. <https://appfinder.orchac.co.uk/> (accessed 7 April 2021).
27. Thornton LK and Kay-Lambkin FJ. Specific features of current and emerging mobile health apps: user views among people with and without mental health problems. *mHealth*; 2018, 4:56, <https://pubmed.ncbi.nlm.nih.gov/30701174/>
28. Peng W, Kanthawala S, Yuan S, et al. A qualitative study of user perceptions of mobile health apps. *BMC Public Health* 2016; 16: 1158.
29. Zapf A, Castell S, Morawietz L, et al. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? *BMC Med Res Methodol* 2016; 16: 93.
30. Krippendorff K. *Content analysis: an introduction to its methodology*. Thousand Oaks: Sage Publications, 1980.
31. Gordon WJ, Landman A, Zhang H, et al. Beyond validation: getting health apps into clinical practice. *Npj Digit Med* 2020; 3: 1–6.
32. Food and Drug Administration. Enforcement Policy for Digital Health Devices For Treating Psychiatric Disorders During the Coronavirus Disease 2019 (COVID-19) Public Health Emergency. 2020, <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enforcement-policy-digital-health-devices-treating-psychiatric-disorders-during-coronavirus-disease>
33. JMIR The model of gamification principles for digital health interventions: evaluation of validity and potential utility | Floryan | Journal of Medical Internet Research. <https://www.jmir.org/2020/6/e16506/> (accessed 18 December 2020).



## Appendix

### Full PsyberGuide Credibility Rating Scale (PGCRS 2.0)

Domain	Feature	Score	Criteria
<b>(1) Intervention Specificity</b>	a. Clarity of proposed goal	2	Product describes at least one mental health goal which is specific, measurable, achievable (e.g. reduce stress, reduce PTSD symptoms)
		1	Product describes non-specific or hard to measure mental health goals (e.g. improve your life, improve your wellbeing)
		0	No clear goals
<b>(2) Consumer Ratings</b>	a. Number/average ratings	2	Ratings exist from >1500 users with an average rating of 3.5+
		1	Ratings exist from 31-1500 users with an average rating of 3.5+
		0	Fewer than 30 user rating OR an average rating below 3.5
<b>(3) Research</b>	a. Direct research evidence	3	Strong research support for the product (at least two between-group design experiments that show efficacy or effectiveness)
		2	Some research support for the product (at least one experiment that shows efficacy or effectiveness)
		1	Other research (e.g. single case designs, quasi-experimental methods demonstrating efficacy, or preliminary analyses)
		0	No research
	b. Indirect research evidence	1	The app uses evidence-based practices to achieve its goals
		0	The app does not use evidence-based practices to achieve its goals (or there are no goals described)
	c. Research Independence & Review	2	At least one research paper funded by government agency (e.g. NIH) or non-profit organization OR two articles published in peer-reviewed journals
		1	All research funded primarily by for-profit organizations or combined funding sources OR one article published in a peer-reviewed journal
0		No information about source of funding for the research AND No published, peer-reviewed papers	

(continued)

Continued.

Domain	Feature	Score	Criteria
<b>(4) Development</b>	a. Development processes	1	Pilot, feasibility and acceptability data OR evidence of stakeholder engagement in development
		0	No pilot, feasibility and acceptability data AND no evidence of stakeholder engagement
	b. Efficacy of Other Products	1	Developer/development team has developed other mental health interventions delivered via technological medium which demonstrate efficacy
		0	No other mental health technological interventions demonstrating efficacy have been developed by this team
	c. Clinical Input in Development	1	Clinical leader with mental health expertise involved in development
		0	No clinical leader with mental health expertise involved in development
	d. Ongoing maintenance & updates	2	The application has been revised within the last 6 months
		1	The application has been revised within the last 12 months
		0	The application has not been revised or was revised more than 12 months ago

**Scoring Instructions**

*For mobile applications:* Assign a score for each feature. Add each feature score to obtain total score. No items need to be reverse coded. To normalize to a 5 point scale, divide total score by 3.

*For web-based tools:* Omit items 2a and 4d. Assign a score for each feature. Add each feature score to obtain total score. No items need to be reverse coded. To normalize to a 5 point scale, multiply total score by  $\frac{5}{11}$ .