Assessing App Quality through Expert Peer Review: A case study from the Gray Matters Study.

Phillip J Hartin, Ian Cleland, Chris D Nugent *Member, IEEE*, Sally I McClean *Member, IEEE*, JoAnn Tschanz, Christine Clark and Maria C Norton.

***Abstract*— Health apps focused on inciting behavior change are becoming increasingly popular. Nevertheless, many lack underlying evidence base, scientific credibility and have limited clinical effectiveness. It is therefore important that apps are well-informed, scientifically credible, peer reviewed and evidence based. This paper presents the use of the Mobile App Rating Scale (MARS) to assess the quality of the Grey Matters app, a cross platform app to deliver health education material and track behavior change across multi-domains with the aim of reducing the risk of developing Alzheimer's disease. The Gray Matters app shows promising results following reviews from 5 Expert raters, achieving a mean overall MARS score of 4.45 ± 0.14. Future work will involve undertaking of a detailed content analysis of behavior change apps to identify common themes and features which may lead to the successful facilitation of sustained behavior change.**

I. Introduction

Health education programs have demonstrated their effectiveness in educating individuals with targeted knowledge relating to risk factors of various diseases [1,2]. With this knowledge, individuals are subsequently capable of making educated decisions regarding lifestyle choices, which may have a significant effect on their future health outcomes. Most health education programs target the leading causes of mortality [3], such as heart disease and stroke [4], cancer [5], diabetes [4] and respiratory diseases [3,6]. Nevertheless, only a limited number of studies have been conducted with a focus on health education for Alzheimer's Disease risk reduction, despite AD and other dementias being the third leading cause of death in the UK in the United Kingdom [7] and the 6th United States [8].

Previous research by the investigators has highlighted the potential of multivariate behavior change interventions to reduce the risk of developing AD [9]. The Gray Matters study was an evidence-based multi-domain lifestyle intervention for middle-aged persons (40 to 64 years) with normal cognition, designed to promote brain health. The six-month RCT of 146 residents of Cache County, Utah (treatment n=104; control n=42) tracked lifestyle behaviors across six domains: physical activity, food choices, social engagement, cognitive stimulation, sleep quality and stress management. Users tracked their physical activity using a wearable activity monitor and self-reported through a smartphone app. Evidence based daily facts, consisting of fact and suggestion pairs, highlighted the link between healthy lifestyle behaviors and improved cognitive wellbeing and were pushed to the user through the app on a daily basis. The aim of the app was to increase knowledge about AD prevention through modifiable lifestyle behaviors and increase intrinsic motivation to change. The primary outcomes of the Gray Matters study were increases in intrinsic motivation, and actual changes in, healthy behaviors, with accompanying reductions in subjective memory complaints. This work has also highlighted the utility of pervasive technologies, such as activity monitors and smart phones, in effective delivery of behavior change interventions. These findings are in agreement with a growing body of supportive literature [10, 11]. Nevertheless, uptake of such interventions has been limited and it is still unclear whether or not the interventions undertaken within research will be sustainable or scalable in a free living environment.

Clearly, there is a wide range of potential use-cases for mobile technology for behavior change within healthcare, nonetheless, the adoption of technology for the purpose of public health education or behavioral change interventions are extremely limited [12, 13]. This may be due to a number of reasons. Firstly, the surge in availability of apps in an unregulated market raises concerns as to the appropriateness quality, inaccurate information/absence of evidence-based content and lack of user and clinician engagement in their development [14, 15].

During the development of the Gray Matters app, the authors undertook a detailed assessment of health related apps in order to insure that both the app and the educational content was of high quality. This paper describes the implementation of a Mobile App Rating Scale (MARS) to assess the quality of the developed application and discusses the use of such a rating scale within a multi domain behavior change application. The Gray matters app is firstly reviewed by Experts. Following this a comparison of The Gray Matters app and those from the original MARS Study [15] is presented.

II. Methods and Procedures

This section provides an overview of the Gray Matters app and describes details of the rating scales and procedures implement to assess app quality.

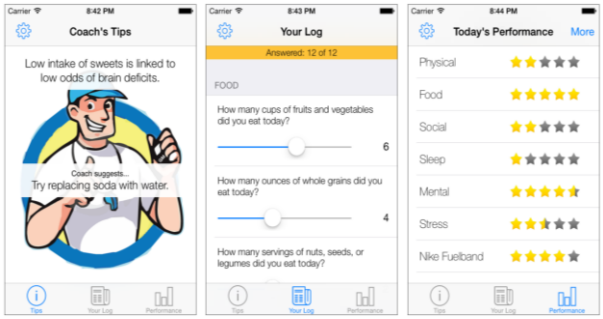
* This work has been supported by the Vice President for Research seed grant, Utah State University and the Department for Employment and Learning, Northern Ireland.

P. J. Hartin, C. D. Nugent and I. Cleland are with the School of Computing and Mathematics, University of Ulster at Jordanstown, Antrim BT37 0QB, U.K. (e-mail: hartin-p1@email.ulster.ac.uk; cd.nugent@ulster.ac.uk; i.cleland@ulster.ac.uk). S. I. McClean is with the School of Computing and Information Engineering, University of Ulster at Coleraine, Londonderry BT52 1SA, U.K. (e-mail: si.mcclean@ulster.ac.uk). J. T. Tschanz, M. C. Norton and C. Clark are with the Department of Psychology, Utah State University, Logan, UT 84322-4440, USA. (e-mail: joann.tschanz@usu.edu, maria.norton@usu.edu, clarkchristinej@gmail.com).

A. *Gray Matters App*

To deliver health education material and track behavior change across the treatment group within the Gray Matters study, a smartphone app (Fig. 1) was developed for iOS and Android [16]. The app facilitated the delivery of health education material through the form of 'factoids'. Each factoid comprised a fact and suggestion pair relating to AD and preventative strategies, e.g. "Low dietary sodium is protective against cognitive de-cline; Use your favorite spice instead of salt to flavor food" (Fig. 1a). In total 164 factoids were produced for the study. A different factoid was delivered by notification, to the participant each morning. To monitor behavioral changes, the participants were requested to self-report their behaviors by answering 12 questions daily (Fig. 1b). Each question related to one of the core domains Physical, Mental, Sleep, Food, Social and stress. Each question had a recommended value, based on the Centre for Disease Control and Prevention's (CDC) minimum daily targets. Using these recommended values, it was possible to provide the participant with immediate feedback as to their efforts in the form of a 5-star rating (refer to Fig. 1c).

Figure 1. Screen shots of the Gray Matters App, a) fact and suggestion pairs, b) selfreporting questios for 6 domains and c)star rating feedback.



(a) (b) (c)

B. *Mobile App Rating Scale (MARS)*

There are a number of methods by which a clinical intervention can be evaluated, including systematic reviews and critical appraisals [17-18]. To evaluate the mobile based technology solution however, options are limited. Stoyanov, Hides, Kavanagh, et al. developed the Mobile App Rating Scale (MARS), which is a peer-reviewed, objective, multidimensional measure for trialing, classifying, and rating the quality of mobile health apps [15]. The scale assesses app quality across 5 core criteria: engagement, functionality, aesthetics, information quality, and subjective quality [15]. Within each criterion, there are a number of sub-items from which to rate (n=23). Each item is graded on a 5-point scale (1-Inadequate to 5-Excellent). The users who rate the developed solution using MARS should be experts within the targeted health domain. A summary score, referred to as the MARS mean score, is calculated as the mean score across the 4 objective criteria (excluding the subjective quality criteria scores). This is the primary measure by which all apps are contrasted or ranked. In the development of the MARS framework, the authors identified 59 mHealth (mobile health) apps that were available to the public via their platforms respective app stores. Of these 59, 9 were randomly selected to develop the scale in a pilot study, and 50 were subsequently used to evaluate the consistency and inter-reliability of the scale. Of the 9 Apps used in the creation of the scale, only 1 addressed more than one behavioural domain. A large proportion of these apps, 66.7%, were primarily aimed at reducing stress, whilst the remainder addressed diet, sleep and social domains. No apps addressed physical activity, diet, or smoking.

C. *Participant Recruitment*

Five expert raters where recruited to review the Gray matters app using the MARS tool. These raters were from 2 relevant disciplines, Medicine (n=4) and Computer Science (n=1). It must be noted that the expert raters in both fields have had professional experience in psychology and/or the study of behavior change. Each rater was asked to download the app and use for 5-10 minutes. After use of the app, the rater was asked to complete the MARS survey. The rater could reuse the app during the completion of the survey to minimize information recall error. To further compare the Gray Matters App, each expert rater was asked to rate an additional 3 apps, two from the original MARS list [15] and 1 from an updated list of more modern apps. The process of how these additional apps are selected is given below.

D. *Selecting Additional Apps.*

As noted earlier, the apps reviewed in the MARS pilot (n=9) were predominately focused on the stress domain. In the full MARS study, an additional 50 apps were included in the assessment of the scale. The original authors disclosed the criteria scores and mean MARS score for each of these 50 apps. These apps were then targeted for inclusion in a comparison with the Gray Matters app. However, initial content analysis of the apps found that a 13.5% (n=8) were not relevant to behavior change in any of the identified domains, or they did not aim to improve health outcomes. As such these 8 were excluded from further study. Of the remaining apps (n=51), the behavioral focus was heavily unbalanced, with 74.5% (n=38) of the apps focusing strictly on the stress domain. In addition, no apps reviewed in the original MARS studies addressed smoking, and only 1 targeted the social domain. As such it was apparent that additional apps should be reviewed to provide a fairer and representative comparison of existing apps in the marketplace.

Identifying suitable apps to compare to the developed solution is difficult. The Gray Matters app is a disease specific app, utilizing education delivery and behavior tracking across numerous behavioral domains. The Gray Matters app is itself a relatively novel concept, thus narrowing the number of potential apps from which to compare. There are however a large number of apps that focus on specific behavioral areas, such as cognitive stimulation, diet monitoring, activity tracking, activity promotion and stress relief. There are also a number of apps that aim to disseminate information on Alzheimer's Disease and other areas of cognitive decline.

Suitable apps for comparison were identified by searching the iOS and Android marketplaces for apps that met one of the following 4 criteria 1. Encouraged behavior change, 2. Aimed to improve health status, 3. Gamified behavior change or 4. Delivered condition specific health education. This

search reviewed over 70 publicly available apps, of which 36 were identified as suitable for comparison. The distribution of primary behavioral domains targeted by these additional apps can be seen in Table 1. The inclusion of these apps helps in some way to balance the distribution of domains targeted from the original study, however, stress remains the most frequently targeted domain (47.1%). All 36 apps were analysed and reviewed by the author.

TABLE I. COMPARISON OF APP DATASETS AND DISTRIBUTION OF PRIMARY BE- HAVIOURAL DOMAIN TARGETED

Domain	MARS		App Market		Consolidated	
	n	%	n	%	n	%
Physical	4	7.80	8	22.20	12	13.80
Diet	2	3.90	7	19.40	9	10.30
Cognitive	2	3.90	7	19.40	9	10.30
Sleep	4	7.80	3	8.30	7	8.00
Social	1	2.00	4	11.10	5	5.70
Stress	38	74.50	3	8.30	41	47.10
Smoking	0	0.00	4	11.10	4	4.60
Total	51	100.00	36	100.00	87	100.00

III. RESULTS AND DISCUSSION

This section presents results from the Expert evaluation. The descriptive statistics of the Gray Matters app ratings can be seen in Table 2. The Gray matters app achieved a mean overall MARS score of 4.45 ± .14. For subjectivity, the app received a mean score of 3.6 ± .379. This is to be expected as the subjectivity score shows the greatest degree of standard deviation, hence it's naming. The app received a true mean score (including Subjective) of 4.28 ± .121. This score is lower than the official MARS score which does not account for the subjective criteria. The difference between the MARS score (M= 4.45, SD = 0.30) and True Mean (M = 4.28, SD = .121) score was found to be statistically significant in a paired samples t-Test at the $p < 0.05$ level [$t(4) = 6.292$, $p = 0.03$].

TABLE II. DESCRIPTIVE STATISTICS OF EXPERT RATERS SCORING FOR EACH ASSESSMENT CATEGORY, AND TOTAL MEANS, INCLUDING AND EXCLUDING SUBJECTIVITY

Criteria	N	Minimum	Maximum	Mean	Std. Deviation
Engagement	5	3.80	4.40	4.1200	.30332
Functionality	5	4.25	4.75	4.5000	.17678
Aesthetics	5	4.67	5.00	4.7360	.14758
Information	5	4.29	4.71	4.4860	.18783
Subjective	5	3.25	4.00	3.6000	.37914
MARS Score	5	4.31	4.60	4.4580	.14307
True Mean	5	4.10	4.47	4.2880	.17513

A. Comparison of Gray matters app with scores from original MARS study

The app scored very highly in all areas. A mean MARS score of 4.45 ranks the Gray Matters app 2nd out of 60 reviewed apps (59 Original + Gray Matters app). When comparing the true mean scores (including subjective score), the app also ranks 2nd. The nearest ranking app in both cases is 'headspace', a structured meditation guide which launched in 2010 [19]. Headspace is marketed as 'a gym membership for the mind', and aims to reduce stress and increase mindfulness for its' users. The app has over 5 million installs on android devices alone. If the MARS score is indicative of public adoption, the resulting Gray Matters app demonstrates the framework's promise as an mHealth development platform. Whilst the results of the expert review are encouraging, steps must be taken to ensure that these ratings were valid and comparable to the original study. To do this, each expert was asked to rate an app from the original MARS study, and from a list of newly sourced apps. In the following section, the Inter-rater Reliability (IRR) and ICC scores for each expert are established.

B. Comparison with additional Apps

Whilst the MARS scale has produced an average rating of 4.45 across 5 expert reviewers, the score exists without true context. It is important to compare this result with that of existing apps, which are aiming for similar outcomes. As such, a number of apps within the area of quantified-self and health improvement were considered for review. To further validate the scores that were attributed by the author, each expert rater, in addition to rating the Gray Matters app, where asked to rate an additional 3 apps, two from the original MARS list [15] and one from the additional 36 apps identified by the author. A random number generator was used to select the app ids. The 3 apps highlighted for review were: PTSD Coach (MARS), Conscious (MARS), Water Your Body (Extra). The scores reached by each expert rater, including the original MARS scores and authors re- evaluation, can be seen in table 3 below.

TABLE III. MARS SCORES ATTRIBUTED BY EACH EXPERT RATER, THE ORIGINAL STUDY, AND THE AUTHOR

App name	MARS	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5
PTSD Coach	4.29	3.1	3.26	3.63	3.1	3.12
Conscious	3.36	2.93	3.38	3.26	3.15	3.3
Water Your Body	N/A	3.7	3.77	4.11	3.96	3.85
Gray Matters	N/A	4.31	4.48	4.59	4.31	4.6

Using the MARS scores from the original study and the scores from the expert reviews, a reliability analysis was performed. Descriptive statistics, detailing means and standard deviations for each rater are displayed in Table 4.

In this instance, we can see that Expert 1 attributes the lowest mean scores across the reviewed apps (n=2), whilst the original MARS score has the highest mean and largest standard deviation. The differences between the MARS score and the experts seems significant. Reliability analysis

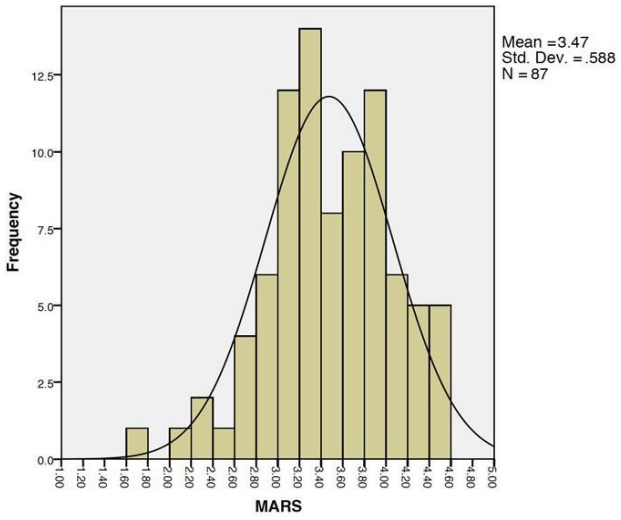
performed using SPSS v22 confirms this, showing an ICC of .167. This suggests poor consistency between the experts and the original MARS study. It appears that the MARS studies results may be the cause of this. To test, reliability analysis was performed between the expert ratings of all apps (n=4), omitting the MARS results. The ICC between the expert raters was found to be .991, displaying excellent consistency.

TABLE IV. EXPERT RATER AND MARS STATISTICS

	Mean	Std. Deviation
Expert 1	3.015	0.12021
Expert 2	3.32	0.08485
Expert 3	3.445	0.26163
Expert 4	3.125	0.03536
Expert 5	3.21	0.12728
MARS	3.825	0.65761

Using the mean expert rating of 4.45, the Gray Matters app is ranked 3rd of the 87 apps reviewed, and is placed in the first decile (>4.29). A histogram displaying the distribution of all 87 apps MARS scores can be seen in Figure 2.

Figure 2. Histogram showing distribution and normal curve of mean MARS scores for all 87 reviewed apps



IV. CONCLUSION

The results from the MARS assessment carried out by expert raters are promising. The reliability found between the experts and the original MARS study however highlight a number of limitations. These are as follows. The number of apps from which the comparisons made were very small. As such, for future comparisons a larger number of apps should be reviewed to find a more representative result. Given that the expert reviewers were predominately from the medical domain; profession or knowledge bias may have affected the results. Future work will aim to include additional experts from varying fields. Furthermore, whilst the apps can be ranked against one another using the MARS score, the reasons/ features that contribute top these scores are not apparent. To understand the factors which influence these rankings a content analysis must be undertaken.

REFERENCES

[1] Mason TA, Thompson WW, Allen D, Rogers D, Gabram-Mendola S, Arriola KRJ. Evaluation of the Avon Foundation community education and outreach initiative Community Patient Navigation Program. Health Promot Pract [Internet] 2013 Jan [cited 2014 Jun 10];14(1):105–12.

[2] Krantz MJ, Coronel SM, Whitley EM, Dale R, Yost J, Estacio RO. Effectiveness of a community health worker cardiovascular risk reduction program in public health and health care settings. Am J Public Health [Internet] 2013 Jan [cited 2014 Jun 10];103(1):e19–27.

[3] Bauer UE, Briss PA, Goodman RA, Bowman BA. Prevention of chronic disease in the 21st century: Elimination of the leading preventable causes of premature death and disability in the USA. Lancet 2014;384:45–52. PMID: 24996589

[4] Centers for Disease Control and Prevention. State Public Health Actions to Prevent and Control Diabetes, Heart Disease, Obesity and Associated Risk Factors and Promote School Health [Internet]. 2015 [cited 2015 Mar 4]. Available from: http://www.webcitation.org/6YKBePuIE

[5] Hocking W. Epidemiologic Studies in Cancer Prevention and Screening [Internet]. 2013. Available from: http://www.springerlink.com/index/10.1007/978-1-4614-5586-8

[6] Pierce JP, White VM, Emery SL. What public health strategies are needed to reduce smoking initiation? Tob Control 2012;21:258–264.

[7] Office for National Statistics. Mortality Statistics: Deaths Registered in England and Wales (Series DR), 2013 [Internet]. London; 2014. Available from: http://www.ons.gov.uk/ons/rel/vsob1/mortality-statistics--deaths-registered-in-england-and-wales--series-dr-/2013/index.html.

[8] Centers for Disease Control and Prevention, Leading Causes of Death Factsheet, [Online] http://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. Accessed: 02/22/16.

[9] Norton, Maria C., et al. "The design and progress of a multidomain lifestyle intervention to improve brain health in middle-aged persons to reduce later Alzheimer's disease risk: The Gray Matters randomized trial." Alzheimer's & Dementia: Translational Research & Clinical Interventions 1.1 (2015):

[10] Kim, Sarang, et al. "Development of the motivation to change lifestyle and health behaviours for dementia risk reduction scale." Dementia and geriatric cognitive disorders extra 4.2 (2014): 172-183.

[11] Kivipelto, Miia, et al. "The Finnish geriatric intervention study to prevent cognitive impairment and disability (FINGER): study design and progress."Alzheimer's & Dementia 9.6 (2013): 657-665.

[12] Lyons, Elizabeth J., et al. "Behavior change techniques implemented in electronic lifestyle activity monitors: a systematic content analysis." Journal of medical Internet research 16.8 (2014).

[13] Topol EJ. The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care. 1st ed. 2012.

[14] Reynoldson, C.; Stones, C.; Allsop, M.; Gardner, P.; Bennett, M. I.; Closs, S. J.; Jones, R.; Knapp, P. Assessing the quality and usability of smartphone apps for pain self-management. Pain Med. 2014, 15, 898–909.

[15] Stoyanov, Stoyan R., et al. "Mobile App Rating Scale: A New Tool for Assessing the Quality of Health Mobile Apps." JMIR mHealth and uHealth 3.1 (2015).

[16] Hartin, P., Nugent, C., McClean, S., *et. al*.: Encouraging Behavioral Change via Everyday Technologies to Reduce Risk of Developing Alzheimer's Disease. In: procedings, L., Chen, L., Nugent, C., and Bravo, J. (eds.) Ambient Assisted Living and Daily Activities, 6th International Work-Conference, 2014 pp. 51–58.

[17] D. L. Sackett, "Evidence-based medicine," Seminars in Perinatology, vol.21, no. 1, pp. 3–5, Feb. 1997, issn: 01460005. doi:10.1016/S0146-0005(97)80013-4. [Online].

[18] J. M. Morrison, F. Sullivan, E. Murray, and B. Jolly, "Evidence-based education: development of an instrument to critically appraise reports of educational interventions," Medical Education, vol. 33, no. 12,pp. 890–893, Dec. 1999, doi: 10.1046/j.1365-2923.1999.00479.x.

[19] HEADSPACE INC., Headspace, 2015. [Online]. Available: https://www.headspace.com (visited on 12/30/2015).