PLOS ONE

ASSESSING MENTAL HEALTH SERVICE USER AND CARER INVOLVEMENT IN PHYSICAL HEALTH CARE PLANNING: THE DEVELOPMENT AND VALIDATION OF A NEW PATIENT-REPORTED OUTCOME MEASURE.

--Manuscript Draft--

Manuscript Number:	PONE-D-18-05375R1
Article Type:	Research Article
Full Title:	ASSESSING MENTAL HEALTH SERVICE USER AND CARER INVOLVEMENT IN PHYSICAL HEALTH CARE PLANNING: THE DEVELOPMENT AND VALIDATION OF A NEW PATIENT-REPORTED OUTCOME MEASURE.
Short Title:	Development and validation of the EQUIP Physical Health PREM
Corresponding Author:	Chris Gibbons Harvard Medical School Boston, MA UNITED KINGDOM
Keywords:	Patient-reported outcomes; PRO; measures; questionnaire; item response theory, computerized adaptive testing; CAT; mokken; analysis; Rasch; mental health; parity of esteem; Schizophrenia; bi-polar disorder; care plan; planning; Patient-Centered Care; patient centred care
Abstract:	BACKGROUND People living with serious mental health conditions experience increased morbidity due to physical health issues driven by medication side-effects and lifestyle factors. Coordinated mental and physical healthcare delivered in accordance with a care plan could help to reduce morbidity and mortality in this population. Efforts to develop new models of care are stymied by a lack of validated instruments to accurately assess the extent to which mental health services users and their carers are involved in care planning for physical health. OBJECTIVE To develop a brief and accurate patient-reported outcome measure (PROM) capable of assessing mental health services user and carer involvement in physical health care planning. METHODS We employed psychometric and statistical techniques to refine a bank of candidate questionnaire items, derived from qualitative interviews, into a valid and reliable measure of service user and carer involvement in care planning for physical health. We assessed the psychometric performance of the item bank using Mokken and Rasch analyses. Our analyses included unidimensionality, scalability, fit to the partial credit Rasch model, category threshold ordering, local dependency, differential item functioning, and test-retest reliability. Once purified of poorly performing and erroneous items, we simulated computerized adaptive tests with 15, 10 and 5 items using the calibrated item bank. RESULTS Issues with category threshold ordering, local dependency and differential item functioning were evident for a number of items in the nascent item bank and were resolved by removing problematic items. The final 19 item PROM had excellent fit to the Rasch model (x2 = 123.58, df = 133, P=.23, RMSEA = .04 (95% CI = 0-.07)) and high reliability (marginal r = 0.93). The correlation between theta scores at baseline and 2-week follow-up was high (r = .70, P < .01) and 94.9% of assessment pairs were within the Bland Altman limits of agreement. Simulated computerized adaptive testing demonstrated that assessments could be made using as few as 10 items (mean SE = .43). DISCUSSION We have developed a flexible patient reported outcome measure to quantify service user and carer involvement in physical health care planning. We demonstrate the potential to substantially reduce assessment length by utilizing computerized adaptive testing administration.
Order of Authors:	Chris Gibbons

	Helen Brooks
	Judith Gellatly
	Nicola Small
	Karina Lovell
	Penny Bee
Opposed Reviewers:	
Response to Reviewers:	Responses to Reviewers for manuscript # PONE-D-18-05375 ASSESSING MENTAL HEALTH SERVICE USER AND CARER INVOLVEMENT IN PHYSICAL HEALTH CARE PLANNING: THE DEVELOPMENT AND VALIDATION OF A NEW PATIENT-REPORTED OUTCOME MEASURE. PLOS ONE To the Editor and Reviewers, We thank the Editorial and Review team for providing an in-depth and useful review. We have been able to respond positively to all of the suggestions made by the team and have detailed the changes which we have made in this document. At the request of the journal, we have added information about our assessment of capacity of consent as well as two anonymized datasets containing the data we used in our research. We believe that the changes have bolstered both the scientific rigor of the work and its likely impact on the field of multidisciplinary care for users and carers involved with serious mental health care services. Kind regards, Chris Gibbons Harvard Medical School Comments to the Author 1. Is the manuscript technically sound, and do the data support the conclusions? The manuscript must describe a technically sound piece of scientific research with data that supports the conclusions. Experiments must have been conducted rigorously, with appropriate controls, replication, and sample sizes. The conclusions must be drawn appropriately based on the data presented. Reviewer #1: Partly Reviewer #2: Yes 2. Has the statistical analysis been performed appropriately and rigorously? Reviewer #1: No Reviewer #2: Yes 3. Have the authors made all data underlying the findings in their manuscript fully available? The PLOS Data policy requires authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception (please refer to the Data Availability Statement in the manuscript PDF file). The data

should be provided as part of the manuscript or its supporting information, or deposited to a public repository. For example, in addition to summary statistics, the data points behind means, medians and variance measures should be available. If there are restrictions on publicly sharing data—e.g. participant privacy or use of data from a third party—those must be specified.

Reviewer #1: Yes

Reviewer #2: Yes

4. Is the manuscript presented in an intelligible fashion and written in standard English?

PLOS ONE does not copyedit accepted manuscripts, so the language in submitted articles must be clear, correct, and unambiguous. Any typographical or grammatical errors should be corrected at revision, so please note any specific errors here.

Reviewer #1: Yes

Reviewer #2: Yes

5. Review Comments to the Author

Please use the space provided to explain your answers to the questions above. You may also include additional comments for the author, including concerns about dual publication, research ethics, or publication ethics. (Please upload your review as an attachment if it exceeds 20,000 characters)

Reviewer #1: Some aspects of the design of the study and the analytical procedures deserve more details or need clarification. In particular:

a) In the selection of the final 19 items from an initial larger pool, the authors used two measurement models (Mokken and Rasch), using the first (Mokken) as a tool for selecting scalable and unidimensional items and the second (Rasch) to refined the selection. I am slightly perplexed by this approach, that, anyway, has to be more adequately motivated in the article. Moreover, the tenure of the assumptions of the Rasch model need to be evaluate and report in the text. This is particularly important with regard to the unidimensionality assumption. Several tools are available, such as Cronbach's Alpha, Principal component analysis conducted on model residuals, the use of the MIRT (Multidimensional Item Response Theory) family models. I urge the Authors to supply information about this aspect.

We thank the Reviewer for their insightful review of our article. We have made a number of additions to the manuscript based on the requests made in this paragraph:

We have added information on the method which we have utilized, especially regarding the interface between Rasch and Mokken analyses and the estimation of dimensionality. We have added a new reference which more clearly states the potential for combining the two methodologies, as well as the existing references which signpost studies that have used the same approach previously. The new text is added below (from Methods, page 4):

“We fitted data from nascent scale to the partial credit “Rasch” model (PCM)13,14 in order to assess psychometric performance. We evaluated factor structure, scalability and monotonicity by fitting data to non-parametric Mokken model before more rigorous psychometric assessments using the PCM.15 The combination of the two methodologies has been shown to be useful in previous research conducted by members of our group and others.16–19
Where scale data did not conform to the assumptions of either the Mokken or the partial credit model, an iterative process of item reduction was undertaken to remove

	the violating items from the analysis.20 The iterative process involved assessments of scalability, model and item fit to the PCM, category threshold disordering, local dependency, and differential item functioning (DIF). Each concept and the method by which it is assessed is described in greater detail below. MOKKEN ANALYSIS The Mokken model is a non-parametric extension of the simple deterministic Guttman scaling model. 21 The model provides a framework to extend the unreastically error-free Guttman models using probabilistic estimation, thus accounting for measurement error.22 As a non-parametric item response theory (NIRT) model, the Mokken models relax some assumptions of item response theory whilst affirming essential assumptions such as unidimensionality and scalability.22 We fitted data to the double monotonicity model, a NIRT model which estimates a single parameter for each item (i.e., the level of the construct which that item assesses). By successfully fitting scale data to a Mokken model it can be said to be both unidimensional and properly scaled. We utilized parallel polychoric principal component analysis which compared the experimental eigenvalues with a Monte Carlo simulated eigenvalues to verify the unidimensional factor structure before proceeding to item response theory analysis.23,24 “ b) After a process of item selection and response categories aggregation very articulated and punctually reported, the Authors arrived at the estimation of a Partial Credit model with 19 items, whose results - however - must be described more in detail. In Table 2, it is not clear what are the parameters "b1" and "b2" (maybe this is a particular notation of the software used?). More importantly, individual item's information related to the SEM (Standard Error of Measurement) and the Fit (e.g.: chi-square statistic; inlier-pattern-sensitive fit statistic or INFIT; outlier-sensitive fit statistic or OUTFIT; etc.) need to be added. In response to these comments we have recreated Table 2. The modified Table, which is copied below here, now includes information on the item fit for each of the items and includes a legend which described the Delta parameters (we have previously used the IRT convention of naming the threshold parameters b1, b1 rather than the Rasch convention of naming them delta1, delta2, etc.) We have also added an item information curve for the entire item bank in a new figure shown below. Figure 4 – Overall Scale Information and Standard Error Key – SE = Standard Error c) It seems (not being explicitly stated in the text) that the number of items from which the analysis began is 62 (or 91?). If so, it is a rather large number of questions, fulfilled by 267 people (most of them seriously ill). Given that the task would be burdensome even for healthy people, it is desirable that the Authors provide some additional information regarding any difficulties encountered during the "data collection" phase (for example: any assistance in completing the questionnaire, possible treatment of responses missing, etc.). We have reported the questionnaire completion rates and provided information on the methodology we used to collect this information. We did not save any data for users or

carers who did not finish the questionnaire. For the responses we did receive, 2966 out of 18291 (16%) responses were missing, models were fitted without imputation although IRT imputation was necessary to calculate model and item fit statistics. The responses are shown below with new text shown in red (page 6).

In the Methods:
“For items flagged as having signficiant DIF following Bonferroni correction, we used the McFadden pseudo R2 estimation with recommend cut-off of R2 > .035 being indicative of meaningful DIF.33 By assessing DIF between service users and carers we will explore the suitablility of the nascent PROM for both groups. Models were fitted with missing data present. However, missing data were imputed using IRT-based estimation.34 Given the well-documented issues with model fit statistics, we prioritized meeting the assumptions of the Mokken and Rasch models over model fit, as has been recommended elsewhere.35”

In the Results:
We collected data from 267 mental health services users from the United Kingdom. 67 participants completed the 67 candidate questionnaire items a second time after two weeks. No data were available on the number of participants who began the survey but did not complete it. Missing responses were given in 2966 or 18291 cells (16%).”

d) Does the final 10-items instrument cover the content domains reported at p. 3? Moreover, in the light of the large number of items showing local dependency (27), how can the Authors exclude that more than 1 latent dimensions were necessary to model the initial pool of items?

The final 19-items cover the themes we uncovered which were relevant to service users and their carers. The exact items which are completed by a user or carer under CAT conditions are difficult to determine without content balancing, which we did not employ in our simulations. We have added a comment in the Discussion section related to this information, which now reads:

“It is noteworthy that when administering CATs each individual respondent is likely to complete different combinations of items which form a subset of the complete item bank. Though the scores between the unidimensional CAT and the fixed-length short-form are highly correlated, there is no guarantee that every patient will complete items from each of the content domains which were nominated by service users and carers. In the current manuscript, we prioritize brevity and accuracy and simulate CAT administration without content balancing or prioritizing certain items. We acknowledge that other users may prioritize item exposure and thus may utilize CATs differently.”

We are not aware of a reason that local dependency ought to signal multidimensionality and so cannot refer to this issue directly. We did employ the AISP procedure in Mokken analysis and have now add an additional stage including a parallel principal components analysis based on a polychoric matrix to demonstrate the sufficiency of extracting a single factor. This change has led to amendments in both the Method and Results sections of the manuscript.

In the Methods:
“We utilized parallel polychoric principal component analysis which compared the experimental eigenvalues with a Monte Carlo simulated eigenvalues to verify the unidimensional factor structure before proceeding to item response theory analysis.22,23”

And in the Results:
“The 57 remaining items were free from violations of monotonicity and were unidimensional. Parallel principal component and factor analysis confirmed the unidimensional structure of the dataset as the eigenvalue for the second factor/component (2.87, 2.16) was below simulated eigenvalues in the Monte Carlo dataset (1.50, 1.19).”
e) In the final paragraph of the article, among the "limits", it is written that: <<... our sample consisted of predominantly white, female service users>>. Perhaps it would be appropriate to clarify that it is not a "sample" (at least in the statistical sense of the

term) and, above all, that - as described in the paragraph "Data Collection" - it is rather a "self-selected" group of people.
We thank the Reviewer for this observation and have corrected the term ‘sample’.

f) Please, add a description of the content of Figures 2, Figure 3 and Table 2 in the text. In general, the presentation of the analysis and results is very compress, more details would be useful for better appreciate the work.

We have added additional descriptions for each Figure and Table and provided detailed legends for Figures 2 and Table 2. On reflection, we decided to remove Figure 3 as it communicated so much information and yet was so difficult to interpret that it did not positively add to the manuscript. The headline figure from that table, the R2 change is still clearly stated in the text.

g) Minor issues
- p. 2, please, indicate what the acronym SUs stands for;
- p. 3, please, indicate what the acronym NRES stands for;
- p. 5, <<...when the probability of a certain response to a question differs across different demographic groups...>>, please, remove “demographic”, also other kinds of grouping variables could be used;
- the Authors mentioned only some of the categories labels used for the 5-point scale items (p. 5). However, a full description of the categories labels is important, especially in reason of the category thresholds analysis they conducted and the resulting decision to aggregate some of the items’ categories.

h) The manuscript need to be read over, several typos are present, for example:
- p. 2, <<...co-morbidities and, , are significantly...>>, please, remove a comma;
- p.2, << The care plan is … with the wishes of both service users and their carers>>, the final period is lacking: carers...
- p. 4, <>, replace “an” with “a”;
- p.4, < .05)>>, delete >.05;
- p.4, <<(see Fig 1..>> replace with (see Fig 1);
- p. 6, replace << RMSEA .097 >> with << RMSEA = .097 >>;
- p.6, Table 1, SU+Carer+SUandC=196+46+33=275; Female+Male =206+62=268, whereas in the title of the table the sample size is 267; We have corrected the Table account for missing data
- p. 6, << A total of 27 individual items that displayed local dependency with more than one other item and were removed from the analysis>>, remove “and”;
- p.7, in the title of Table 2, scale parameters are mentioned, but they did not appear in the table;
- p. 7, the item showing a Dif was item 65, as indicated in the text, or item 22 as indicated in Figure 3?
- p. 8, Table 2, item 7, 9 and 16: the wording seems not complete, moreover, uniform the use of periods;
- p. 8, at the beginning of Computer Adaptive Testing paragraph, text character need to be uniformed.

We have addressed all of the minor issues which the Reviewer has helpfully highlighted. We thank the Reviewer for their careful reading of our manuscript.

Reviewer #2: The paper by Gibbons deals with an interesting and very actual topic, that is the user involvement in physical health care planning. Physical health of people with severe mental disorders has been recognized as one of the most cogent problems in the field of mental health, given the significant reduced life expectancy of patients with severe mental disorders. Some issues need to be addressed in a revision of the paper:

•The introduction section is not well organized, and paragraphs seem not linked each other. Moreover, this section lacks a clear focus on the main issue of the paper, that is the development of a new assessment questionnaire to promote the participation of users in physical health care planning. I would suggest the author to rewrite this section.
•In several sections of the paper (including the title) the importance of the involvement of carers on health care planning is emphasized, but in the study only patients have

been involved. This is a little bit confusing. I would suggest to remove the terms “carers” through the text, unless they have been included in the study (and therefore this should be clarified).

We thank the Reviewer for their observation. We have made substantial amendments to the Introduction paragraph. We believe that the amended section is clearer and has better flow.

We have clarified in the manuscript that carers were involved and that the final PROM is suitable for both users and carers. For example (new text in italics):

Methods
“For items flagged as having signficiant DIF following Bonferroni correction, we used the McFadden pseudo R2 estimation with recommend cut-off of R2 > .035 being indicative of meaningful DIF.29 By assessing DIF between service users and carers we will explore the suitablility of the nascent PROM for both groups.”

Discussion
“The new PROM contains 19 items which were successfully fitted to a single-parameter Rasch item response theory model. The PROM is suitable for assessing both service users and carers. The 19 items also serve as an item bank for computerized adaptive testing…”

•The authors mention that several decisions regarding patients’ health are usually taken without an active involvement of patients, and that a shared decision making should be the basis of any patient-clinician interaction. This issue should be expanded since it represents one of the main aims of the paper. Authors could quote papers coming from the EU funded CEDAR and ROAMER projects, as examples of good collaboration with users and carers in mental health.

We thank the Reviewer for this excellent suggestion. We have added relevant references to outputs from the both the CEDAR and ROAMER initiatives. For example (next text in italics):

“Despite research consistently showing that this sort of involvement is aligned with the desires of both service users and carers there is a paucity of care models which have been shown to effectively increase involvement in this way.7 More broadly, increasing the quality of mental health services was the top research priority expressed by an international working group comprising both professionals as well as users and carers.8”

•The method section is too poor. In order to improve this section, I would recommend to report: 1) a detailed description of the development of the initial version of the questionnaire (i.e., were focus groups organized in order to define the most important issues to be included in the questionnaire? Was a literature search performed in order to evaluate the state of the art? Were experts in the field interviewed?); 2) the recruitment process (i.e., how and when patients will be recruited, inclusion and exclusion criteria).

We have made substantial amendments to the method section which over the following points raised by the Reviewer, including description of the study in which the data were originally collected and greater details on the psychometric analyses.

•In the Results’ section, I would suggest providing some descriptive data regarding the global sample. I would strongly recommend to provide information on the clinical characteristics of patients, such as duration of illness, severity of symptoms, presence of comorbidities (e.g., hypertension, diabetes, hypertriglyceridemia, etc.), pharmacological treatments (not only including psychotropic drugs), smoking habits, physical activity (or physical inactivity). If these data are not available, this represents a major bias of the study which should be acknowledged.

We utilized a recruitment approach which was designed to capture a broad demographic of service users and carers in the United Kingdom. We are pleased that this approach allowed us to gain better geographic representation (which we now display in Table 1) than we were likely to get with in-clinic recruitment alone. The disadvantage of our methodology, as the Reviewer points out, is that we cannot accurately assess the presence of comorbidities, treatments, or substances. Our experience in conducting this sort of research is that self-report of this information is both off-putting to patients and unlikely to be of high quality.

We acknowledge this limitation in a statement in the Discussion which reads (new text italicized):

“The current study has some limitations. Firstly, our dataset consisted of predominantly white, female service users. Though all systematic differences between demographic groups were corrected for in the current analysis, further research would be warranted to ensure that the items perform well in groups which were not well represented in our data. It should be noted that whilst we demonstrated uniform scale performance across demographic groups – including service users and carers, we did not collect information relating to comorbidities, physical activity or substance and further research would be necessary to explicitly confirm that the scale is unaffected by differences in disease or lifestyle factors within groups of service users.”

•Moreover, the authors should avoid to repeat the type of statistical test adopted several times in the text, since it has been already stated in the relevant paragraph.

We have reduced the instances in which we repeat the name of the statistical test which we used.

•The Discussion section is very brief and not useful. A comparison with available literature is needed. This section should be organized as follows: a) a brief summary of main findings of the study; b) a comparison with previous available studies; 3) strengths and limitations.

We have amended the Discussion section along as the Reviewer has suggested. We have added sections to make the manuscript more useful, such as a section on how CAT may be implemented. We have made more references to the previous literature as well as explicitly discussing the strengths of our work. We make reference to the other relevant initiatives which the Reviewer has suggested.

The entirely-novel sections are copied below and further amendments are presented in the attached manuscript:
“The measure will facilitate benchmarking of service quality and service user experience, aligned with contemporary philosophies and policies for collaborative recovery-focused mental health care. The philosophy of the new PREM is that mental and physical health are equally important (the so-called parity of esteem), and parity of esteem is increasingly being embedded in policy and practice imperatives derived from stakeholder consultation.46”

“Parties who wish to use CAT administration for the EQUIP-PH measure are directed towards many packages available for the R Statistical Programming Environment including mirt and catR.34,47 One tool for implementing CATs is the Concerto platform, developed and maintained by the University of Cambridge.48 Further details can be found on the Concerto website (concertoplatform.com) or by request to the authors of this manuscript.”

“Our study also has some notable strengths. We have collected a geographically diverse group of both service users and carers and created a flexible assessment which can be used without modification of assessing and comparing both groups. The EQUIP-PH PREM which we have developed is related to the EQUIP measure, a questionnaire measure for service user and carer involvement in care planning, which was recently developed by our group9. Both tools could be used together to gain a holistic understanding of how involved service users and carers are in mental health

	care planning. Further research could usefully be conducted to understand the scores from the two instruments in relation to one another and provide further insight into their use as a tool to assess global care planning and service delivery. “ •A conclusive paragraph should be useful, in which the importance of this study for the clinical practice should be highlighted. “In conclusion, The EQUIP-PH PROM is a brief, accurate, and flexible service user- and carer-reported assessment for involvement in physical health care planning for users of serious mental health services. The measure provides a reliable means to evaluate and benchmark the quality of physical health management in the context of mental health care.” •The text should be revised: there are many misprints throughout the text. We have carefully revised the text and amended the misprints. 6. If you would like your identity to be revealed to the authors, please include your name here (optional). Your name and review will not be published with the manuscript. Reviewer #1: (No Response) Reviewer #2: (No Response) [NOTE: If reviewer comments were submitted as an attachment file, they will be attached to this email and accessible via the submission site. Please log into your account, locate the manuscript record, and check for the action link "View Attachments". If this link does not appear, there are no attachment files to be viewed.] While revising your submission, please upload your figure files to the Preflight Analysis and Conversion Engine (PACE) digital diagnostic tool, http://pace.apexcovantage.com/. PACE helps ensure that figures meet PLOS requirements. To use PACE, you must first register as a user. Registration is free. Then, login and navigate to the UPLOAD tab, where you will find detailed instructions on how to use the tool. If you encounter any issues or have any questions when using PACE, please email us at figures@plos.org. Please note that Supporting Information files do not need this step. --
Additional Information:	
Question	**Response**
Financial Disclosure Please describe all sources of funding that have supported your work. **This information is required for submission and will be published with your article, should**	This research was funded by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care (NIHR CLAHRC) Greater Manchester. The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

it be accepted. A complete funding statement should do the following: 1. Include **grant numbers and the URLs** of any funder's website. Use the full name, not acronyms, of funding institutions, and use initials to identify authors who received the funding. 2. **Describe the role** of any sponsors or funders in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. If the funders had **no role** in any of the above, include this sentence at the end of your statement: "*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript*." However, if the study was **unfunded**, please provide a statement that clearly indicates this, for example: "*The author(s) received no specific funding for this work*." * typeset	
Competing Interests You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or non-financial competing interests. Do any authors of this manuscript have competing interests (as described in the PLOS Policy on Declaration and Evaluation of Competing Interests)? **If yes**, please provide details about any and all competing interests in the box	The authors have all declared that no competing interests exist.

below. Your response should begin with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests:* **If no** authors have any competing interests to declare, please enter this statement in the box: "*The authors have declared that no competing interests exist.*" * typeset	
Ethics Statement You must provide an ethics statement if your study involved human participants, specimens or tissue samples, or vertebrate animals, embryos or tissues. All information entered here should **also be included in the Methods section** of your manuscript. Please write "N/A" if your study does not require an ethics statement. **Human Subject Research (involved human participants and/or tissue)** All research involving human participants must have been approved by the authors' Institutional Review Board (IRB) or an equivalent committee, and all clinical investigation must have been conducted according to the principles expressed in the Declaration of Helsinki. Informed consent, written or oral, should also have been obtained from the participants. If no consent was given, the reason must be explained (e.g. the data were analyzed anonymously) and reported. The form of consent (written/oral), or reason for lack of consent, should be indicated in the Methods section of your manuscript.	The study and all associated procedures were approved by the London - West London and Gene Therapy Advisory Committee (GTAC) Research Ethics Committee (16/LO/0386) in February 2016.

Please enter the name of the IRB or Ethics Committee that approved this study in the space below. Include the approval number and/or a statement indicating approval of this research. **Animal Research (involved vertebrate animals, embryos or tissues)** All animal work must have been conducted according to relevant national and international guidelines. If your study involved non-human primates, you must provide details regarding animal welfare and steps taken to ameliorate suffering; this is in accordance with the recommendations of the Weatherall report, "The use of non-human primates in research." The relevant guidelines followed and the committee that approved the study should be identified in the ethics statement. If anesthesia, euthanasia or any kind of animal sacrifice is part of the study, please include briefly in your statement which substances and/or methods were applied. Please enter the name of your Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board, and indicate whether they approved this research or granted a formal waiver of ethical approval. Also include an approval number if one was obtained. **Field Permit** Please indicate the name of the institution or the relevant body that granted permission.	
Data Availability	No - some restrictions will apply

PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the PLOS Data Policy and FAQ for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found. Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. **Please note that simply stating 'data available on request from the author' is not acceptable.** ***If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.*** Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	
Please describe where your data may be found, writing in full sentences. **Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted.** If you are copying our sample text below, please ensure you replace any instances of XXX with the appropriate details. • If your data are all contained within the paper and/or Supporting Information files, please state this in your answer below. For example, "All relevant data are within the paper and its Supporting Information files." • If your data are held or will be held in a public repository, include URLs, accession numbers or DOIs. For example, "All XXX files are available from the XXX database (accession	Anonymized data is available from the corresponding author upon request.

number(s) XXX, XXX)." If this information will only be available after acceptance, please indicate this by ticking the box below. • If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in the box below. For example: “Data are available from the XXX Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data.” “Data are from the XXX study whose authors may be contacted at XXX.” * typeset	
Additional data availability information:	Tick here if your circumstances are not covered by the questions above and you need the journal’s help to make your data available.

HARVARD MEDICAL SCHOOL

Brigham and Women's Hospital

75 FRANCIS STREET
BOSTON, MA 02115

Faulkner Hospital

1153 CENTRE STREET, SUITE 21
BOSTON, MA 02130

To the Editor and Reviewers,

We thank the Editorial and Review team for providing an in-depth and useful review. We have been able to respond positively to all of the suggestions made by the team. We have added anonymized datasets containing the information we used in our analyses.

We believe that the changes have bolstered both the scientific rigor of the work and its likely impact on the field of multidisciplinary care for users and carers involved with serious mental health care services.

Details of our response are copied with this letter.

We look forward to your response,

Kind regards,

Dr. Chris Sidey-Gibbons Ph.D. CPsychol AFBPsS
Co-Director, Patient-Reported Outcomes, Value & Experience (PROVE) Center
Member of the Faculty of Surgery, Harvard Medical School

Responses to Reviewers for manuscript # PONE-D-18-05375
ASSESSING MENTAL HEALTH SERVICE USER AND CARER INVOLVEMENT IN PHYSICAL HEALTH CARE PLANNING: THE DEVELOPMENT AND VALIDATION OF A NEW PATIENT-REPORTED OUTCOME MEASURE.
PLOS ONE

Comments to the Author

1. Is the manuscript technically sound, and do the data support the conclusions?

The manuscript must describe a technically sound piece of scientific research with data that supports the conclusions. Experiments must have been conducted rigorously, with appropriate controls, replication, and sample sizes. The conclusions must be drawn appropriately based on the data presented.

Reviewer #1: Partly

Reviewer #2: Yes

---

2. Has the statistical analysis been performed appropriately and rigorously?

Reviewer #1: No

Reviewer #2: Yes

---

3. Have the authors made all data underlying the findings in their manuscript fully available?

The PLOS Data policy requires authors to make all data underlying the findings described in their manuscript fully available without restriction, with rare exception (please refer to the Data Availability Statement in the manuscript PDF file). The data should be provided as part of the manuscript or its supporting information, or deposited to a public repository. For example, in addition to summary statistics, the data points behind means, medians and variance measures should be available. If there are restrictions on publicly sharing data—e.g. participant privacy or use of data from a third party—those must be specified.

Reviewer #1: Yes

Reviewer #2: Yes

---

4. Is the manuscript presented in an intelligible fashion and written in standard English?

PLOS ONE does not copyedit accepted manuscripts, so the language in submitted articles must be clear, correct, and unambiguous. Any typographical or grammatical errors should be corrected at revision, so please note any specific errors here.

Reviewer #1: Yes

Reviewer #2: Yes

---

5. Review Comments to the Author

Please use the space provided to explain your answers to the questions above. You may also include additional comments for the author, including concerns about dual publication, research ethics, or publication ethics. (Please upload your review as an attachment if it exceeds 20,000 characters)

Reviewer #1: Some aspects of the design of the study and the analytical procedures deserve more details or need clarification. In particular:

a) In the selection of the final 19 items from an initial larger pool, the authors used two measurement models (Mokken and Rasch), using the first (Mokken) as a tool for selecting scalable and unidimensional items and the second (Rasch) to refined the selection. I am slightly perplexed by this approach, that, anyway, has to be more adequately motivated in the article. Moreover, the tenure of the assumptions of the Rasch model need to be evaluate and report in the text. This is particularly important with regard to the unidimensionality assumption. Several tools are available, such as Cronbach's Alpha, Principal component analysis conducted on model residuals, the use of the MIRT (Multidimensional Item Response Theory) family models. I urge the Authors to supply information about this aspect.

We thank the Reviewer for their insightful review of our article. We have made a number of additions to the manuscript based on the requests made in this paragraph:

We have added information on the method which we have utilized, especially regarding the interface between Rasch and Mokken analyses and the estimation of dimensionality. We have added a new reference which more clearly states the potential for combining the two methodologies, as well as the existing references which signpost studies that have used the same approach previously. The new text is added below (from Methods, page 4):

"We fitted data from nascent scale to the partial credit "Rasch" model (PCM)[13,14] in order to assess psychometric performance. We evaluated factor structure, scalability and monotonicity by fitting data to non-parametric Mokken model before more rigorous psychometric assessments using the PCM.[15] The combination of the two methodologies has been shown to be useful in previous research conducted by members of our group and others.[16–19]
Where scale data did not conform to the assumptions of either the Mokken or the partial credit model, an iterative process of item reduction was undertaken to remove the violating items from the analysis.[20]
The iterative process involved assessments of scalability, model and item fit to the PCM, category threshold disordering, local dependency, and differential item functioning (DIF). Each concept and the method by which it is assessed is described in greater detail below.

MOKKEN ANALYSIS
The Mokken model is a non-parametric extension of the simple deterministic Guttman scaling model. [21] The model provides a framework to extend the unreastically error-free Guttman models using probabilistic estimation, thus accounting for measurement error.[22] As a non-parametric item response theory (NIRT) model, the Mokken models relax some assumptions of item response theory whilst affirming essential assumptions such as unidimensionality and scalability.[22] We fitted data to the double monotonicity model, a NIRT model which estimates a single parameter for each item (*i.e.,* the level of the construct which that item assesses). By successfully fitting scale data to a Mokken model it can be said to be both unidimensional and properly scaled. We utilized parallel polychoric principal component analysis which compared the experimental eigenvalues with a Monte Carlo simulated eigenvalues to verify the unidimensional factor structure before proceeding to item response theory analysis.[23,24]
"

b) After a process of item selection and response categories aggregation very articulated and punctually reported, the Authors arrived at the estimation of a Partial Credit model with 19 items, whose results - however - must be described more in detail. In Table 2, it is not clear what are the parameters "b1" and "b2" (maybe this is a particular notation of the software used?). More importantly, individual item's information related to the SEM (Standard Error of Measurement) and the Fit (e.g.: chi-square statistic; inlier-pattern-sensitive fit statistic or INFIT; outlier-sensitive fit statistic or OUTFIT; etc.) need to be added.

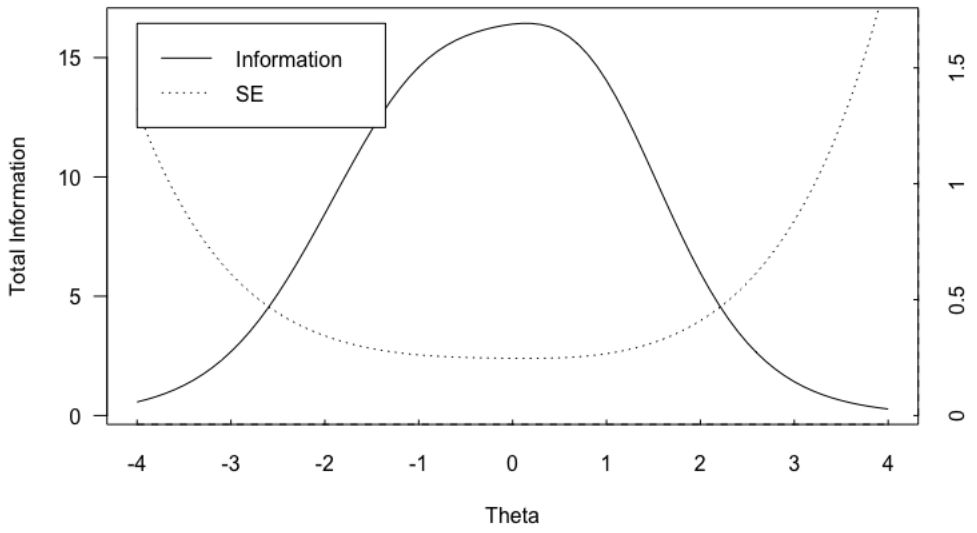
In response to these comments we have recreated Table 2. The modified Table, which is copied below here, now includes information on the item fit for each of the items and includes a legend which described the Delta parameters (we have previously used the IRT convention of naming the threshold parameters b1, b1 rather than the Rasch convention of naming them delta1, delta2, etc.)

We have also added an item information curve for the entire item bank in a new figure shown below.

Item Number	Original number	Wording	Model fit			Item Threshold Parameters		item fit statistics			
			χ^2	df	P	delta 1	delta 2	χ^2	df	P	Scoring
1	50	My care planning team ask about my existing physical health conditions.	21.81	15	.11	-1.20	.12	30.8	34	0.63	0-1-1-2-2
2	24	The physical health information in my care plan is personalised.	13.29	17	.72	-1.92	.20	38.07	34	0.29	0-1-1-2-2
3	53	My care planning team encourage me to take responsibility for my physical health care planning.	14.10	16	.59	-2.01	.27	24.53	32	0.82	0-1-1-2-2
4	37	My opinion on my physical health is valued by my care planning team.	13.48	16	.64	-2.03	.38	26.98	31	0.67	0-1-1-2-2
5	4	I know who reads the physical health information contained within my care plan	11.79	15	.69	-1.02	.43	27.49	33	0.74	0-1-1-2-2
6	55	My care planning team offer practical advice about my physical health.	15.44	15	.42	-1.42	.51	33.18	31	0.36	0-1-1-2-2
7	13	My care plan gives details of my physical health history.	20.20	15	.16	-1.16	.52	26.99	31	0.67	0-1-1-2-2
8	15	My thoughts about my physical health are included in my care plan.	11.20	17	.85	-.62	.76	28.12	29	0.51	0-1-1-2-2
9	52	I experience continuity of care for the treatment of both my physical health conditions and mental health conditions.	15.22	17	.58	-.68	.77	39.73	32	0.16	0-1-1-2-2
10	22	The physical health information in my care plan is helpful.	9.43	19	.97	-1.61	.83	28.01	32	0.67	0-1-1-2-2
11	16	Physical health reviews are carried out in a timely manner.	14.55	16	.56	-.53	.86	28.02	32	0.67	0-1-1-2-2
12	62	My care planning team have a good understanding of my fears about future physical health conditions.	15.30	18	.64	-.77	.92	28.4	32	0.65	0-1-1-2-2
13	56	My care planning team have the time they need to talk to me about physical health concerns.	28.39	19	.08	-1.18	.94	27.15	28	0.51	0-1-1-2-2
14	44	The content of my physical health care plan is responsive to changes in my circumstances.	15.06	19	.72	-1.14	1.06	26.01	33	0.8	0-1-1-2-2
15	27	Information in my care plan has helped me to maintain my physical health.	13.12	18	.78	-.63	1.08	26.09	33	0.8	0-1-1-2-2
16	41	I was asked what I wanted in the physical health information in my care plan.	14.07	16	.59	-.25	1.13	32.37	34	0.55	0-1-1-2-2
17	60	The care plan adequately addresses any side effects I experience from my medication.	14.79	19	.74	-1.22	1.13	40.21	35	0.25	0-1-1-2-2
18	46	I have had the opportunity to invite all the relevant people to care planning meetings related to my physical health.	17.23	16	.37	-.55	1.27	31.57	33	0.54	0-1-1-2-2
19	51	My care planning team makes sure my mental health is not prioritised over my physical health.	21.01	19	.34	-1.15	1.51	31.72	32	0.48	0-1-1-2-2

Figure 4 – Overall Scale Information and Standard Error

Key – SE = Standard Error



c) It seems (not being explicitly stated in the text) that the number of items from which the analysis began is 62 (or 91?). If so, it is a rather large number of questions, fulfilled by 267 people (most of them seriously ill). Given that the task would be burdensome even for healthy people, it is desirable that the Authors provide some additional information regarding any difficulties encountered during the "data collection" phase (for example: any assistance in completing the questionnaire, possible treatment of responses missing, etc.).

We have reported the questionnaire completion rates and provided information on the methodology we used to collect this information. We did not save any data for users or carers who did not finish the questionnaire. For the responses we did receive, 2966 out of 18291 (16%) responses were missing, models were fitted without imputation although IRT imputation was necessary to calculate model and item fit statistics. The responses are shown below with new text shown in red (page 6).

In the Methods:

"For items flagged as having signficiant DIF following Bonferroni correction, we used the McFadden pseudo R2 estimation with recommend cut-off of R2 > .035 being indicative of meaningful DIF.[33] By assessing DIF between service users and carers we will explore the suitablility of the nascent PROM for both groups. Models were fitted with missing data present. However, missing data were imputed using IRT-based estimation.[34] Given the well-documented issues with model fit statistics, we prioritized meeting the assumptions of the Mokken and Rasch models over model fit, as has been recommended elsewhere.[35]"

In the Results:

We collected data from 267 mental health services users from the United Kingdom. 67 participants completed the 67 candidate questionnaire items a second time after two weeks. No data were available on the number of participants who began the survey but did not complete it. Missing responses were given in 2966 or 18291 cells (16%)."

d) Does the final 10-items instrument cover the content domains reported at p. 3? Moreover, in the light of the large number of items showing local dependency (27), how can the Authors exclude that more than 1 latent dimensions were necessary to model the initial pool of items?

The final 19-items cover the themes we uncovered which were relevant to service users and their carers. The exact items which are completed by a user or carer under CAT conditions are difficult to determine without content balancing, which we did not employ in our simulations. We have added a comment in the Discussion section related to this information, which now reads:

"It is noteworthy that when administering CATs each individual respondent is likely to complete different combinations of items which form a subset of the complete item bank. Though the scores between the unidimensional CAT and the fixed-length short-form are highly correlated, there is no guarantee that every patient will complete items from each of the content domains which were nominated by service users and carers. In the current manuscript, we prioritize brevity and accuracy and simulate CAT administration without content balancing or prioritizing certain items. We acknowledge that other users may prioritize item exposure and thus may utilize CATs differently."

We are not aware of a reason that local dependency ought to signal multidimensionality and so cannot refer to this issue directly. We did employ the AISP procedure in Mokken analysis and have now add an additional stage including a parallel principal components analysis based on a polychoric matrix to demonstrate the sufficiency of extracting a single factor. This change has led to amendments in both the Method and Results sections of the manuscript.

In the Methods:
"We utilized parallel polychoric principal component analysis which compared the experimental eigenvalues with a Monte Carlo simulated eigenvalues to verify the unidimensional factor structure before proceeding to item response theory analysis.[22,23]"

And in the Results:

"The 57 remaining items were free from violations of monotonicity and were unidimensional. *Parallel principal component and factor analysis confirmed the unidimensional structure of the dataset as the eigenvalue for the second factor/component (2.87, 2.16) was below simulated eigenvalues in the Monte Carlo dataset (1.50, 1.19)."*

e) In the final paragraph of the article, among the "limits", it is written that: <<... our sample consisted of predominantly white, female service users>>. Perhaps it would be appropriate to clarify that it is not a "sample" (at least in the statistical sense of the term) and, above all, that - as described in the paragraph "Data Collection" - it is rather a "self-selected" group of people.

We thank the Reviewer for this observation and have corrected the term 'sample'.

f) Please, add a description of the content of Figures 2, Figure 3 and Table 2 in the text. In general, the presentation of the analysis and results is very compress, more details would be useful for better appreciate the work.

We have added additional descriptions for each Figure and Table and provided detailed legends for Figures 2 and Table 2. On reflection, we decided to remove Figure 3 as it communicated so much information and yet was so difficult to interpret that it did not positively add to the manuscript. The headline figure from that table, the R^2 change is still clearly stated in the text.

g) Minor issues
- p. 2, please, indicate what the acronym SUs stands for;
- p. 3, please, indicate what the acronym NRES stands for;
- p. 5, <<...when the probability of a certain response to a question differs across different demographic groups...>>, please, remove "demographic", also other kinds of grouping variables could be used;
- the Authors mentioned only some of the categories labels used for the 5-point scale items (p. 5). However, a full description of the categories labels is important, especially in reason of the category thresholds analysis they conducted and the resulting decision to aggregate some of the items' categories.

h) The manuscript need to be read over, several typos are present, for example:
- p. 2, <<...co-morbidities and, , are significantly...>>, please, remove a comma;
- p.2, << The care plan is … with the wishes of both service users and their carers>>, the final period is lacking: carers...
- p. 4, <>, replace "an" with "a";
- p.4, < .05)>>, delete >.05;
- p.4, <<(see Fig 1..>> replace with (see Fig 1);
- p. 6, replace << RMSEA .097 >> with << RMSEA = .097 >>;
- p.6, Table 1, SU+Carer+SUandC=196+46+33=275; Female+Male =206+62=268, whereas in the title of the table the sample size is 267; We have corrected the Table account for missing data

- p. 6, << A total of 27 individual items that displayed local dependency with more than one other item and were removed from the analysis>>, remove “and”;
- p.7, in the title of Table 2, scale parameters are mentioned, but they did not appear in the table;
- p. 7, the item showing a Dif was item 65, as indicated in the text, or item 22 as indicated in Figure 3?
- p. 8, Table 2, item 7, 9 and 16: the wording seems not complete, moreover, uniform the use of periods;
- p. 8, at the beginning of Computer Adaptive Testing paragraph, text character need to be uniformed.

We have addressed all of the minor issues which the Reviewer has helpfully highlighted. We thank the Reviewer for their careful reading of our manuscript.

Reviewer #2: The paper by Gibbons deals with an interesting and very actual topic, that is the user involvement in physical health care planning. Physical health of people with severe mental disorders has been recognized as one of the most cogent problems in the field of mental health, given the significant reduced life expectancy of patients with severe mental disorders. Some issues need to be addressed in a revision of the paper:

•The introduction section is not well organized, and paragraphs seem not linked each other. Moreover, this section lacks a clear focus on the main issue of the paper, that is the development of a new assessment questionnaire to promote the participation of users in physical health care planning. I would suggest the author to rewrite this section.
•In several sections of the paper (including the title) the importance of the involvement of carers on health care planning is emphasized, but in the study only patients have been involved. This is a little bit confusing. I would suggest to remove the terms “carers” through the text, unless they have been included in the study (and therefore this should be clarified).

We thank the Reviewer for their observation. We have made substantial amendments to the Introduction paragraph. We believe that the amended section is clearer and has better flow.

We have clarified in the manuscript that carers were involved and that the final PROM is suitable for both users and carers. For example (new text in italics):

Methods
“For items flagged as having signficiant DIF following Bonferroni correction, we used the McFadden pseudo R2 estimation with recommend cut-off of R2 > .035 being indicative of meaningful DIF.[29] *By assessing DIF between service users and carers we will explore the suitablility of the nascent PROM for both groups*.”

Discussion
“The new PROM contains 19 items which were successfully fitted to a single-parameter Rasch item response theory model. *The PROM is suitable for assessing both service users and carers*. The 19 items also serve as an item bank for computerized adaptive testing…”

•The authors mention that several decisions regarding patients’ health are usually taken without an active involvement of patients, and that a shared decision making should be the basis of any patient-clinician interaction. This issue should be expanded since it represents one of the main aims of the paper. Authors could quote papers coming from the EU funded CEDAR and ROAMER projects, as examples of good collaboration with users and carers in mental health.

We thank the Reviewer for this excellent suggestion. We have added relevant references to outputs from the both the CEDAR and ROAMER initiatives. For example (next text in italics):

“Despite research consistently showing that this sort of involvement is aligned with the desires of both service users and carers there is a paucity of care models which have been shown to effectively increase involvement in this way.[7] *More broadly, increasing the quality of mental health services was the top research priority expressed by an international working group comprising both professionals as well as users and carers.*[8]”

•The method section is too poor. In order to improve this section, I would recommend to report: 1) a detailed description of the development of the initial version of the questionnaire (i.e., were focus groups organized in order to define the most important issues to be included in the questionnaire? Was a literature search performed in order to evaluate the state of the art? Were experts in the field interviewed?); 2) the recruitment process (i.e., how and when patients will be recruited, inclusion and exclusion criteria).

We have made substantial amendments to the method section which over the following points raised by the Reviewer, including description of the study in which the data were originally collected and greater details on the psychometric analyses.

•In the Results' section, I would suggest providing some descriptive data regarding the global sample. I would strongly recommend to provide information on the clinical characteristics of patients, such as duration of illness, severity of symptoms, presence of comorbidities (e.g., hypertension, diabetes, hypertriglyceridemia, etc.), pharmacological treatments (not only including psychotropic drugs), smoking habits, physical activity (or physical inactivity). If these data are not available, this represents a major bias of the study which should be acknowledged.

We utilized a recruitment approach which was designed to capture a broad demographic of service users and carers in the United Kingdom. We are pleased that this approach allowed us to gain better geographic representation (which we now display in Table 1) than we were likely to get with in-clinic recruitment alone. The disadvantage of our methodology, as the Reviewer points out, is that we cannot accurately assess the presence of comorbidities, treatments, or substances. Our experience in conducting this sort of research is that self-report of this information is both off-putting to patients and unlikely to be of high quality.

We acknowledge this limitation in a statement in the Discussion which reads (new text italicized):

"The current study has some limitations. Firstly, our dataset consisted of predominantly white, female service users. *Though all systematic differences between demographic groups were corrected for in the current analysis, further research would be warranted to ensure that the items perform well in groups which were not well represented in our data. It should be noted that whilst we demonstrated uniform scale performance across demographic groups – including service users and carers, we did not collect information relating to comorbidities, physical activity or substance and further research would be necessary to explicitly confirm that the scale is unaffected by differences in disease or lifestyle factors within groups of service users."*

•Moreover, the authors should avoid to repeat the type of statistical test adopted several times in the text, since it has been already stated in the relevant paragraph.

We have reduced the instances in which we repeat the name of the statistical test which we used.

•The Discussion section is very brief and not useful. A comparison with available literature is needed. This section should be organized as follows: a) a brief summary of main findings of the study; b) a comparison with previous available studies; 3) strengths and limitations.

We have amended the Discussion section along as the Reviewer has suggested. We have added sections to make the manuscript more useful, such as a section on how CAT may be implemented. We have made more references to the previous literature as well as explicitly discussing the strengths of our work. We make reference to the other relevant initiatives which the Reviewer has suggested.

The entirely-novel sections are copied below and further amendments are presented in the attached manuscript:

"The measure will facilitate benchmarking of service quality and service user experience, aligned with contemporary philosophies and policies for collaborative recovery-focused mental health care. The philosophy of the new PREM is that mental and physical health are equally important (the so-called parity of esteem), and parity of esteem is increasingly being embedded in policy and practice imperatives derived from stakeholder consultation.[46]''

"Parties who wish to use CAT administration for the EQUIP-PH measure are directed towards many packages available for the R Statistical Programming Environment including mirt and catR.[34,47] One tool for implementing CATs is the Concerto platform, developed and maintained by the University of Cambridge.[48] Further details can be found on the Concerto website (concertoplatform.com) or by request to the authors of this manuscript.''

"Our study also has some notable strengths. We have collected a geographically diverse group of both service users and carers and created a flexible assessment which can be used without modification of assessing and comparing both groups. The EQUIP-PH PREM which we have developed is related to the EQUIP measure, a questionnaire measure for service user and carer involvement in care planning, which was recently developed by our group[9]. Both tools could be used together to gain a holistic understanding of how involved service users and carers are in mental health care planning. Further research could usefully be conducted to understand the scores from the two instruments in relation to one another and provide further insight into their use as a tool to assess global care planning and service delivery. "

•A conclusive paragraph should be useful, in which the importance of this study for the clinical practice should be highlighted.

"In conclusion, The EQUIP-PH PROM is a brief, accurate, and flexible service user- and carer-reported assessment for involvement in physical health care planning for users of serious mental health services. The measure provides a reliable means to evaluate and benchmark the quality of physical health management in the context of mental health care."

•The text should be revised: there are many misprints throughout the text.

We have carefully revised the text and amended the misprints.

---

6. If you would like your identity to be revealed to the authors, please include your name here (optional).

Your name and review will not be published with the manuscript.

Reviewer #1: (No Response)

Reviewer #2: (No Response)

[NOTE: If reviewer comments were submitted as an attachment file, they will be attached to this email and accessible via the submission site. Please log into your account, locate the manuscript record, and check for the action link "View Attachments". If this link does not appear, there are no attachment files to be viewed.]

While revising your submission, please upload your figure files to the Preflight Analysis and Conversion Engine (PACE) digital diagnostic tool, http://pace.apexcovantage.com/. PACE helps ensure that figures meet PLOS requirements. To use PACE, you must first register as a user. Registration is free. Then, login and navigate to the UPLOAD tab, where you will find detailed instructions on how to use the tool. If you encounter any issues or have any questions when using PACE, please email us at figures@plos.org. Please note that Supporting Information files do not need this step.

--

ASSESSING MENTAL HEALTH SERVICE USER AND CARER INVOLVEMENT IN PHYSICAL HEALTH CARE PLANNING: THE DEVELOPMENT AND VALIDATION OF A NEW PATIENT-REPORTED EXPERIENCE MEASURE.

Sidey-Gibbons CJ[1,2], Brooks H[3], Gellatly J[4], Small N[5], Lovell K[4], Bee P[4*]

1. Patient-reported Outcomes, Value, and Experience (PROVE) Center, Brigham and Women's Hospital, Boston, MA
2. Department of Surgery, Harvard Medical School, Boston MA.
3. Department of Psychological Sciences, Institute of Psychology, Health and Society, University of Liverpool, Liverpool, UK
4. Mental Health Research Group, Division of Nursing, Midwifery and Social Work, Faculty of Biology, Medicine and Health, School of Health Sciences, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK
5. NIHR School of Primary Care Research, Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

*= Corresponding author.
e-mail = penny.bee@manchester.ac.uk

All authors contributed equally to this work.

ABSTRACT

BACKGROUND

People living with serious mental health conditions experience increased morbidity due to physical health issues driven by medication side-effects and lifestyle factors. Coordinated mental and physical healthcare delivered in accordance with a care plan could help to reduce morbidity and mortality in this population. Efforts to develop new models of care are hampered by a lack of validated instruments to accurately assess the extent to which mental health services users and carers are involved in care planning for physical health.

OBJECTIVE

To develop a brief and accurate patient-reported experience measure (PREM) capable of assessing involvement in physical health care planning for mental health service users and their carers.

METHODS

We employed psychometric and statistical techniques to refine a bank of candidate questionnaire items, derived from qualitative interviews, into a valid and reliable measure involvement in physical health care planning. We assessed the psychometric performance of

the item bank using modern psychometric analyses. We assessed unidimensionality, scalability, fit to the partial credit Rasch model, category threshold ordering, local dependency, differential item functioning, and test-retest reliability. Once purified of poorly performing and erroneous items, we simulated computerized adaptive testing (CAT) with 15, 10 and 5 items using the calibrated item bank.

Results

Issues with category threshold ordering, local dependency and differential item functioning were evident for a number of items in the nascent item bank and were resolved by removing problematic items. The final 19 item PREM had excellent fit to the Rasch model fit (x^2 = 192.94, df = 1515, *P*=.02, RMSEA = .03 (95% CI = .01-.04). The 19-item bank had excellent reliability (marginal r = 0.87). The correlation between questionnaire scores at baseline and 2-week follow-up was high (r = .70, *P* < .01) and 94.9% of assessment pairs were within the Bland Altman limits of agreement. Simulated CAT demonstrated that assessments could be made using as few as 10 items (mean SE = .43).

Discussion

We developed a flexible patient reported outcome measure to quantify service user and carer involvement in physical health care planning. We demonstrate the potential to substantially reduce assessment length whilst maintaining reliability by utilizing CAT.

Introduction

People diagnosed with severe mental illnesses, such as disorder schizophrenia and bipolar disorders, exhibit higher rates of physical co-morbidities and, as a result, are significantly more likely to die prematurely than the general population.[1–3]

Factors contributing to this deterioration in physical health for mental health service users are known to include side-effects from anti-psychotic medications, higher rates of smoking and substance abuse, poor nutrition, and physical inactivity.[4] Though the relationship between serious mental health issues, physical cormorbidity, and reduced life expectancy is well understood, far less is known about how to organize care delivery to improve physical health and reduce the risk of associated morbidity in this population. Recent evidence suggests that, despite increased awareness of these issues, mortality risk associated with all mental health conditions is rising internationally.[2]

One approach to improve the management of known risk factors is individualized care planning; [5,6] an approach which involves service users and carers working collaboratively with professionals to co-develop a written care plan. This plan aims to accurately document the core issues that a service user would like to address as part of their mental health recovery.

A growing body of research shows that, although collaborative care planning is s aligned with the desires of both service users and carers there is a paucity of care models which

have been shown to effectively increase involvement in care planning for physical health in this way.[7] More broadly, increasing the quality of mental health services was the top research priority expressed by an international working group comprising both professionals as well as users and carers.[8]

Progress in the development of interventions to improve care planning involvement between service users, carers, and providers is stymied by the lack of a meaningful outcome assessment. Quantification of abstract subjective phenomena, such as involvement with care planning, is best accomplished by directly assessing the perspective of the service user or carer; usually using a tool commonly referred to as a patient-reported outcome measure (PREM).

Patient reported outcome measures are a efficient and accurate way to quantify the views of service users and their carers. A relevant example is the EQUIP PREM, which was developed by our group to assess service user and carer involvement in mental health care planning.[9] Previous research has highlighted the importance of brief assessments for mental health service users and their carers, with a strong user preference for minimising response burden by developing shorter questionnaires.[9,10] New assessment modalities including computerized adaptive testing (CAT), are able to tailor person-centred assessments to the individual, a process which tends to result in shorter, more relevant assessments. [11]

The objective of the current paper is to create a novel PREM to assess mental health service user and carer involvement in physical health care planning. We seek to develop a PREM that is accurate, reliable, and suitable for individualized CAT assessment.

METHODS

ITEM DEVELOPMENT METHODS

A set of 67 candidate items were developed following qualitative interviews with mental health service users (SUs), their carers, and mental health professionals from the UK. Further details of the qualitative interview process can be found in a separate manuscript.[12] Items were developed to reflect six pre-identified themes; three of which covered general mental health care planning requirements and three of which were unique to physical health care planning. The general themes included: tailoring a collaborative working relationship between the service users and their carers and the service providers, maintaining a trusting relationship with a professional, having access to a tangible document which could be edited and updated. The physical health themes were: valuing physical health equally with mental health, experiencing coordinated care between health professionals in different disciplines, and having a personalised physical heath discussion.

DATA COLLECTION

Potential participants who expressed an interest in taking part in the study were given a participant information sheet written to current UK National Research Ethics Service (NRES) guidelines. We worked with service users and carers to co-develop the information sheet. The information sheet included details on the study including the potential risks and benefits of taking part, the ways in which participants could take part in the study (e.g.

online via SelectSurvey or through the completion and return of paper versions of the questionnaire), and provided potential participants with the contact details of researchers should they wish to discuss their involvement prior to taking part. All participants responded affirmatively to the question "I have read and understood the participant information sheet" and consent was implied by the completion and return of questionnaires. The study and all associated procedures were approved by the London – West London and Gene Therapy Advisory Committee (GTAC) Research Ethics Committee (16/LO/0386) in February 2016.

Data analysis

We fitted data from nascent scale to the partial credit "Rasch" model (PCM)[13,14] in order to assess psychometric performance. We evaluated factor structure, scalability and monotonicity by fitting data to non-parametric Mokken model before more rigorous psychometric assessments using the PCM.[15] The combination of the two methodologies has been shown to be useful in previous research conducted by members of our group and others.[16–19] Where scale data did not conform to the assumptions of either the Mokken or the partial credit model, an iterative process of item reduction was undertaken to remove the violating items from the analysis.[20] The iterative process involved assessments of scalability, model and item fit to the PCM, category threshold disordering, local dependency, and differential item functioning (DIF). Each concept and the method by which it is assessed is described in greater detail below.

Mokken Analysis

The Mokken model is a non-parametric extension of the simple deterministic Guttman scaling model. [21] The model provides a framework to extend the unreastically error-free Guttman models using probabilistic estimation, thus accounting for measurement error.[22] As a non-parametric item response theory (NIRT) model, the Mokken models relax some assumptions of item response theory whilst affirming essential assumptions such as unidimensionality and scalability.[22] We fitted data to the double monotonicity model, a NIRT model which estimates a single parameter for each item (*i.e.,* the level of the construct which that item assesses). By successfully fitting scale data to a Mokken model it can be said to be both unidimensional and properly scaled. We utilized parallel polychoric principal component analysis which compared the experimental eigenvalues with a Monte Carlo simulated eigenvalues to verify the unidimensional factor structure before proceeding to item response theory analysis.[23,24]

The Partial Credit Model

The PCM is a measurement model which describes the probabalistic relationship between the assessment and the respondent as an interaction between the amount of the latent construct that the respondent has (*i.e.* involvement with physical health care planning) and the level of the latent construct which the item measures. Both the amount of the construct that the respondent has and the level of the latent construct that the item measures can be described in terms of theta (θ). For example, a item which measures a very high level of physical health care planning (which would be an question that we would not expect many people to affirm; for example the questionnaire item may ask about service user or carer's access to a document containing a detailed strategy for physical health care) would be less likely to be affirmed than an item measuring a low level of the latent construct (which would

be a question that we would expect many people to affirm; for example the questionnaire item may ask about whether a health care professional had asked whether a service user was receiving any care for physical health issues).

Goodness-of-fit statistics can be used to assess the data's fit to the PCM model at both the item and scale level. In this study we used both the Chi-square and root-mean square error of approximation to evaluate the fit to the model. We accepted both a non-significant Chi-square interaction ($P > .05$) and RMSEA ($< .05$) indicating good fit. [25]

Category Threshold Ordering

In the case of a Likert or 'multiple choice' item response the probability of responding to each category is modelled separately. As the level of the underlying construct (*i.e.,* involvement in physical health care planning) rises the probability of responding to each Likert category rises to a peak before falling. Different probabilities are given for each response category at every level of θ. It is essential that each category becomes the most likely response option at a certain level of θ. If this is not the case the item is said to exhibit category threshold disordering.

Category threshold disordering refers to the situation in which one or more of the Likert response categories are not the most likely response at any point of along the underlying θ continuum. In the case of disordered category thresholds, we 'collapsed' adjacent categories so that they received the same score. Care was taken not to collapse categories if it were semantically illogical to do so, (*i.e.*, "Agree" would not be collapsed into "Neither Agree nor Disagree"). An illustrated side-by-side example is provided in a previous paper from our group. [26]

Item response theory models (of which the Rasch model is a special case) are predicated on the assumption that differences in responses to items are driven solely by changes in the underlying trait.[27] One way in which items can violate this assumption is local dependency, a situation whereby the response to one item is dependent on the response to another.[28] In practice, this can occur where items are too similar. Local dependency is assessed using Yen's Q3 statistic, in which the correlation of item residuals are compared, and item pairs with residual correlations beyond a threshold are said to be locally dependent. We set the threshold to be equal to .2 + the average observed residual.[29,30]

Local dependency can be resolved in a number of ways, including subtesting (where locally dependent items are joined into a 'super' item) and item deletion.[31] As we began with a large bank of candidate items, we elected to remove items which were locally dependant. Our strategy was to remove an item if it were locally dependent with more than one other item and then, in the case that a locally independent item pair only demonstrated dependency with one another, item information curves for each item were examined alongside the item wording and the item which provided less information was removed from further analysis.

Another issue which can interfere with the assumption that differences in item scores ought to be driven solely by differences in the underlying trait is differential item functioning (DIF).[32] Differential item functioning occurs when the probability of a certain response to a

question varies across different demographic groups. For example, if men were more likely to respond affirmatively to a certain item than women despite having an equal level of overall involvment with physical health care planning, that question would be said to be affected by DIF. We used the iterative hybrid ordinal logistic regression/item response theory approach to conduct DIF analyses. For items flagged as having signficiant DIF following Bonferroni correction, we used the McFadden pseudo R2 estimation with recommend cut-off of R2 > .035 being indicative of meaningful DIF.[33] By assessing DIF between service users and carers we will explore the suitablility of the nascent PREM for both groups. Models were fitted with missing data present. However, missing data were imputed using IRT-based estimation.[34] Given the well-documented issues with model fit statistics, we prioritized meeting the assumptions of the Mokken and Rasch models over model fit, as has been recommended elsewhere.[35]

Items that violated any of the above assumptions were removed, and the remaining items were reanalyzed. We evaluated the reliability of the final scale and the overall fit to the Rasch model. Once a final set of purified questions were calibrated by fitting them to the PCM, we simulated computerized adaptive tests (CATs)[36]. The CAT algorithms conducted stepwise assessments by iteratively selecting the item which will maximise the test information based on the participant based on participant's θ estimate which is based off their previous responses. The first item for the assessment is the item which maximises information at the mean population level of θ.

We simulated CATs to assess the viability of brief assessment using the nascent scale using the final items as a 'bank' of candidate items. In computerized adaptive testing an algorithm is used to select the next most appropriate item for the patients based on their previous responses. This approach has been shown to substantially reduce the length of tick-box assessments whilst maintaining, and even increasing, reliability.[18,26]

We similated CATs using the Firestar script for R. Firestar uses a Bayesian expected a posteriori θ estimator and selected items based on the maximum posterior weighted information (MPWI) criterion. The MPWI selects items based on the item information weighted by the posterior distribution of trait/phenomena values.[37]This criterion has been shown to provide excellent measurement information for CAT using polytomous items.

SOFTWARE

Analyses were conducted using the R Statistical Computing Environment with the 'mokken', 'mirt', 'lordif', 'psych', 'ggplot2', 'methcomp' and 'BlandAltLeh' packages installed. [34,38–42]. Computer adaptive testing simulation was conducted using the FIRESTAR script, which was modified to add additional statistics.[43] This modified FIRESTAR code is available on request from the authors.

RESULTS

We collected data from 267 mental health services users from the United Kingdom. 67 participants completed the 67 candidate questionnaire items a second time after two weeks. No data were available on the number of participants who began the survey but did not complete it. 16% of PREM data was missing.

Table 1 Demographic information for 267 participants recruited into the study

Age	44(14)
Gender	69.3% Female
	21% Male
	9.7% unreported
Ethnicity	
	83.7% White
	16.3% Non-white
Service user/carer status	
SU	66%
Carer	15%
Both SU and carer	12%
Not reported	7%
Geographic location	
Northern England	32%
The Midlands	23.7%
Southern England	39.4%
Ireland	1%
Scotland	1.5%
Wales	2.5%

Key: SU = Service users

MOKKEN ANALYSIS

Mokken analysis revealed violations of monotonicity for a number of items (5, 6, 8, 10, 25, 26, 40). In addition, Loevinger's scalability coefficient was too low (Item H >.30) for items 7, 9, 39. The 57 remaining items were free from violations of monotonicity and were unidimensional. Parallel principal component and factor analysis confirmed the unidimensional structure of the dataset as the eigenvalue for the second factor/component (2.87, 2.16) was below simulated eigenvalues in the Monte Carlo dataset (1.50, 1.19).

RASCH ANALYSIS

The remaining items were fitted to the partial credit model. The initial fit to the model was poor (RMSEA > .10); thus prompting evaluation of item performance in the context of Rasch model assumptions. There appeared to be substantial threshold disordering throughout the scale. A single solution was chosen to rescore all items (0-1-1-2-2). The amended threshold probability curves for all the items can be see in Appendix 1.

Model fit improved slightly after rescoring but was still unacceptable (RMSEA = .097 (95% CI = .091-.10)). We evaluated the correlations between item residuals, which were above the threshold in a number of instances. In total, 96 item pairs were locally dependant (see Appendix 2). A total of 27 individual items that displayed local dependency with more than one other item were removed from the analysis. Four sets of items remained which were locally dependant with one another. The item information curves for both pairs of locally-dependent items were compared side-by-side (see Fig 2) and in each case the item with the lowest item information removed from the scale.

Figure 1. Comparison of item information curves for locally dependent items

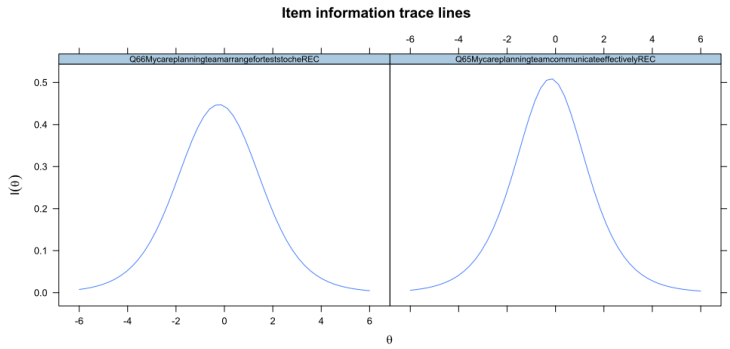


Figure 2. Shows item information curves for a pair of locally-dependent items. The amount of information which each gives about the participant is given on the *y*-axis and the level of underlying construct that the person has (i.e., involvement with physical health care planning) is on the *x*-axis.

No DIF was detected for age or gender but item 65 “My care planning team communicates effectively” was found to have significant DIF (R2 change = .06) between service users and carers.

Following adjustment for category threshold ordering, local dependency, and differential item functioning; item 36 “I feel comfortable attending discussions about my care plan” misfit the model and was removed (x^2 = 34.81, df = 15, *P*=.003). The removal of item 36 led to a final item bank of 19 items which were free from breaches of assumptions of the Rasch model, displayed excellent model fit (x^2 = 192.94, df = 1515, *P*=.02, RMSEA = .03 (95% CI = .01-.04). The 19-item bank had excellent reliability (marginal r = 0.87).

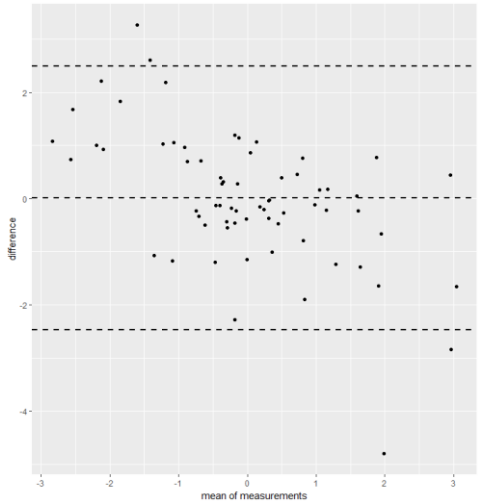
Table 2 Details of final items and item threshold parameters

Item Number	Original number	Wording	Model fit			Item Threshold Parameters		item fit statistics			
			χ^2	df	P	delta 1	delta 2	χ^2	df	P	Scoring
1	50	My care planning team ask about my existing physical health conditions.	21.81	15	.11	-1.20	.12	30.8	34	0.63	0-1-1-2-2
2	24	The physical health information in my care plan is personalised.	13.29	17	.72	-1.92	.20	38.07	34	0.29	0-1-1-2-2
3	53	My care planning team encourage me to take responsibility for my physical health care planning.	14.10	16	.59	-2.01	.27	24.53	32	0.82	0-1-1-2-2
4	37	My opinion on my physical health is valued by my care planning team.	13.48	16	.64	-2.03	.38	26.98	31	0.67	0-1-1-2-2
5	4	I know who reads the physical health information contained within my care plan	11.79	15	.69	-1.02	.43	27.49	33	0.74	0-1-1-2-2
6	55	My care planning team offer practical advice about my physical health.	15.44	15	.42	-1.42	.51	33.18	31	0.36	0-1-1-2-2
7	13	My care plan gives details of my physical health history.	20.20	15	.16	-1.16	.52	26.99	31	0.67	0-1-1-2-2
8	15	My thoughts about my physical health are included in my care plan.	11.20	17	.85	-.62	.76	28.12	29	0.51	0-1-1-2-2
9	52	I experience continuity of care for the treatment of both my physical health conditions and mental health conditions.	15.22	17	.58	-.68	.77	39.73	32	0.16	0-1-1-2-2
10	22	The physical health information in my care plan is helpful.	9.43	19	.97	-1.61	.83	28.01	32	0.67	0-1-1-2-2
11	16	Physical health reviews are carried out in a timely manner.	14.55	16	.56	-.53	.86	28.02	32	0.67	0-1-1-2-2
12	62	My care planning team have a good understanding of my fears about future physical health conditions.	15.30	18	.64	-.77	.92	28.4	32	0.65	0-1-1-2-2
13	56	My care planning team have the time they need to talk to me about physical health concerns.	28.39	19	.08	-1.18	.94	27.15	28	0.51	0-1-1-2-2
14	44	The content of my physical health care plan is responsive to changes in my circumstances.	15.06	19	.72	-1.14	1.06	26.01	33	0.8	0-1-1-2-2
15	27	Information in my care plan has helped me to maintain my physical health.	13.12	18	.78	-.63	1.08	26.09	33	0.8	0-1-1-2-2
16	41	I was asked what I wanted in the physical health information in my care plan.	14.07	16	.59	-.25	1.13	32.37	34	0.55	0-1-1-2-2
17	60	The care plan adequately addresses any side effects I experience from my medication.	14.79	19	.74	-1.22	1.13	40.21	35	0.25	0-1-1-2-2
18	46	I have had the opportunity to invite all the relevant people to care planning meetings related to my physical health.	17.23	16	.37	-.55	1.27	31.57	33	0.54	0-1-1-2-2
19	51	My care planning team makes sure my mental health is not prioritised over my physical health.	21.01	19	.34	-1.15	1.51	31.72	32	0.48	0-1-1-2-2

TEST-RETEST RELIABILITY

The correlation between theta scores at baseline and 2-week follow-up was high (r = .70, *P* < .01). Bland Altman analysis revealed that 94.9% of assessment pairs were within the 95% limits of agreement (see Figure 2).

Figure 2 – Bland Altman plot for test-retest reliability



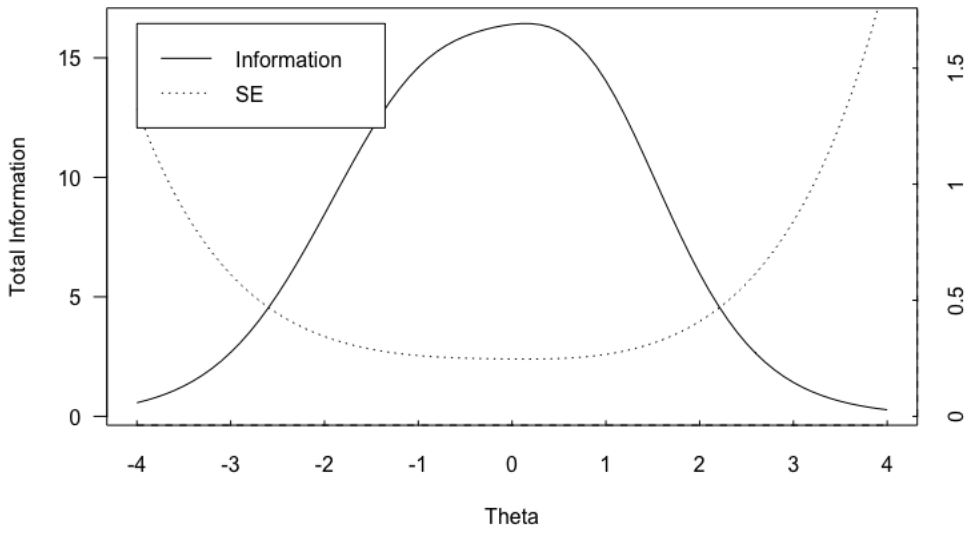
COMPUTERIZED ADAPTIVE TESTING

Adaptive testing simulations were conducted with a simulated Gaussian N (-0.08,1.90) distribution, which matched the distribution of the data used to develop the item banks. Results of CAT simulation are shown in the Table 3. Assessments as short as 10 items demonstrated high correlation with the total score of the full scale. The overall information and standard error which was available in the entire item banks is displayed in Figure 3.

Table 3. Summary of simulated computer adaptive tests.

	Standard Error (SE)			
Number of items	**Mean**	**SD**	**Range**	**Correlation with full scale**
19	.36	.04	.32-.56	1
15	.41	.04	.38-.60	.98
10	.43	.04	.40-.65	.95
5	.66	.04	.62-.78	.87

Figure 3 – Overall Scale Information and Standard Error
Key – SE = Standard Error



DISCUSSION

We present the co-development and validation of a new service user and carer-reported assessment for physical health care planning in serious mental health services, the EQUIP Physical Health PREM (EQUIP-PH-PREM).

The new PREM contains 19 items which were successfully fitted to a single-parameter Rasch item response theory model. The PREM is suitable for assessing both service users and carers. The 19 items also serve as an item bank for computerized adaptive testing (CAT) which can tailor assessments to the individuals who complete the PREM. We show that using CAT administration could substantially reduce burden of response by reducing the number of items in the assessment from 19 to 10, whilst still maintaining acceptable accuracy and high correlation between scores.

In the EQUIP-PH-PREM we provide a tool to support investigations into the experience of service users and carers who are receiving care for a severe mental illness from mental health providers. Adequate service user and carer involvement in care planning decisions are predicated on successful interaction both within and between stakeholder groups. In order to ensure the new PREM incorporated these important aspects the items were developed in collaboration with service users, carers and mental health professionals.

The final PREM items include those that cover having the opportunity and time to be able to discuss physical health concerns, reflecting previously identified organisational barriers to providing integrated care.[12] Similarly, they highlight the importance of *co-created* care plans, which are known to be highly valued by both service users and carers [44,45] Further items serve to facilitate long-term self-management skills that are required to manage physical health concerns.

This new PREM has operationalized the evidence-based best practice framework developed previously which will allow health care providers and service users to challenge current practice by quantifying service user and carer involvement from the user perspective.[12]

The measure will facilitate benchmarking of service quality and service user experience, aligned with contemporary philosophies and policies for collaborative recovery-focused mental health care. The philosophy of the new PREM is that mental and physical health are equally important (the so-called parity of esteem), and parity of esteem is increasingly being embedded in policy and practice imperatives derived from stakeholder consultation.[46]

The EQUIP-PH PREM assesses issues which have been consistently highlighted in consultations with service users and carers and, as such, is well suited for use as an tool to assess the outcome of interventions. Other relevant interventions include those designed to improve inter- and intra- professional communication including professional training and improved health systems to enhance the integration and continuity of care for those under the care of health services.

The current study has some limitations. Firstly, our dataset consisted of predominantly white, female service users. Though all systematic differences between demographic groups were corrected for in the current analysis, further research would be warranted to ensure that the items perform well in groups which were not well represented in our data. It should be noted that whilst we demonstrated uniform scale performance across demographic groups – including service users and carers, we did not collect information relating to comorbidities, physical activity or substance and further research would be necessary to explicitly confirm that the scale is unaffected by differences in disease or lifestyle factors within groups of service users.

Our study is also limited by the necessity to evaluate the CATs using simulated, rather than actual, data. This technique is likely to the slightly over-estimate the accuracy of the CAT as it does not take into account aberrant responders who do not conform to the expectations of the model. Our previous research developing item banks for depression and quality of life suggests that this effect is marginal and that CAT assessment is efficient and precise

both when simulations are made using participant data and when the CAT is deployed in the real world.[18]

It is noteworthy that when administering CATs each individual respondent is likely to complete different combinations of items which form a subset of the complete item bank. Though the scores between the unidimensional CAT and the fixed-length short-form are highly correlated, there is no guarantee that every patient will complete items from each of the content domains which were nominated by service users and carers. In the current manuscript, we prioritize brevity and accuracy and simulate CAT administration without content balancing or prioritizing certain items. We acknowledge that other users may prioritize item exposure and thus may utilize CATs differently.

Parties who wish to use CAT administration for the EQUIP-PH measure are directed towards many packages available for the R Statistical Programming Environment including mirt and catR.[34,47] One tool for implementing CATs is the Concerto platform, developed and maintained by the University of Cambridge.[48] Further details can be found on the Concerto website (concertoplatform.com) or by request to the authors of this manuscript.

Our study also has some notable strengths. We have collected a geographically diverse group of both service users and carers and created a flexible assessment which can be used without modification of assessing and comparing both groups. The EQUIP-PH PREM which we have developed is related to the EQUIP measure, a questionnaire measure for service user and carer involvement in care planning, which was recently developed by our group[9]. Both tools could be used together to gain a holistic understanding of how involved service users and carers are in mental health care planning. Further research could usefully be conducted to understand the scores from the two instruments in relation to one another and provide further insight into their use as a tool to assess global care planning and service delivery.

In conclusion, The EQUIP-PH PREM is a brief, accurate, and flexible service user- and carer-reported assessment of involvement in physical health care planning for users of mental health services with serious mental illnesses. The measure provides a reliable means to evaluate and benchmark the quality of physical health management in the context of mental health care.

Acknowledgement

This research was funded by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care (NIHR CLAHRC) Greater Manchester. The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supporting information.

Supporting Information 1 : Anonymized baseline dataset
Supporting Information 2 : Anonymized follow up dataset

References

1. Harris EC, Barraclough B. Excess mortality of mental disorder. *Br J Psychiatry*. 1998;173(1):11-53. doi:10.1192/BJP.173.1.11.
2. Walker ER, McGee RE, Druss BG. Mortality in Mental Disorders and Global Disease Burden Implications. *JAMA Psychiatry*. 2015;72(4):334. doi:10.1001/jamapsychiatry.2014.2502.
3. Rodgers M, Dalton J, Harden M, Street A, Parker G. Integrated care to address the physical health needs of people with severe mental illness: a rapid review. 2016. https://www.ncbi.nlm.nih.gov/books/NBK355962/. Accessed December 18, 2017.
4. Brown S, Birtwistle J, Roe L, Thompson C. The unhealthy lifestyle of people with schizophrenia. *Psychol Med*. 1999;29(3):697-701. doi:10.1017/S0033291798008186.
5. Doyle C, Lennox L, open DB-B, 2013 undefined. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *bmjopen.bmj.com*. http://bmjopen.bmj.com/content/3/1/e001570.short. Accessed December 18, 2017.
6. Care Quality Commission. *Survey of Mental Health Inpatient Services*.; 2009.
7. Bee P, Brooks H, Fraser C, Lovell K. Professional perspectives on service user and carer involvement in mental health care planning: a qualitative study. *Int J Nurs Stud*. 2015;52(12):1834-1835. http://www.sciencedirect.com/science/article/pii/S0020748915002308. Accessed December 18, 2017.
8. Fiorillo A, Luciano M, Del Vecchio V, Sampogna G, Obradors-Tarragó C, Maj M. Priorities for mental health research in Europe: A survey among national stakeholders' associations within the ROAMER project. *World Psychiatry*. 2013;12(2):165-170. doi:10.1002/wps.20052.
9. Bee P, Gibbons C, Callaghan P, Fraser C, Lovell K. Evaluating and Quantifying User and Carer Involvement in Mental Health Care Planning (EQUIP): Co-Development of a New Patient-Reported Outcome Measure. *PLoS One*. 2016;11(3):e0149973. doi:10.1371/journal.pone.0149973.
10. Gibbons CJ, Bee PE, Walker L, Price O, Lovell K. Service user- and carer-reported measures of involvement in mental health care planning: methodological quality and acceptability to users. *Front psychiatry*. 2014;5:178. doi:10.3389/fpsyt.2014.00178.
11. Gibbons C, Bower P, Lovell K, Valderas J, Skevington S. Electronic quality of life assessment using computer-adaptive testing. *J Med Internet Res*. 2016;18(9):e240.
12. Small N, Brooks H, Grundy A, et al. Understanding experiences of and preferences for service user and carer involvement in physical health care discussions within mental health care planning. *BMC Psychiatry*. 2017;17(1):138. doi:10.1186/s12888-017-1287-1.
13. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research; 1960.
14. Masters GN. A rasch model for partial credit scoring. *Psychometrika*. 1982;47(2):149-174. doi:10.1007/BF02296272.
15. Mokken RJ. *A Theory and Procedure of Scale Analysis: With Applications in Political Research*. Walter de Gruyter; 1971. https://books.google.com/books?hl=en&lr=&id=vAumIrkzYj8C&pgis=1. Accessed February 16, 2015.

16. Meijer RR, Sijtsma K, Smid NG. Theoretical and Empirical Comparison of the Mokken and the Rasch Approach to IRT. *Appl Psychol Meas*. 1990;14(3):283-298. doi:10.1177/014662169001400306.
17. Stochl J, Jones PB, Croudace TJ. Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC Med Res Methodol*. 2012;12(1):74. doi:10.1186/1471-2288-12-74.
18. Loe BS, Stillwell D, Gibbons C. Computerized Adaptive Testing Provides Reliable and Efficient Depression Measurement Using the CES-D Scale. *J Med Internet Res*. 2017;19(9):e302. doi:10.2196/jmir.7453.
19. Gibbons CJ, Small N, Rick J, Burt J, Hann M, Bower P. The Patient Assessment of Chronic Illness Care produces measurements along a single dimension: results from a Mokken analysis. *Health Qual Life Outcomes*. 2017;15(1):61. doi:10.1186/s12955-017-0638-4.
20. Pallant J, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol*. 2007;46(1):1-18.
21. Guttman L. The basis for scalogram analysis. 1949. https://scholar.google.co.uk/scholar?hl=en&q=The+basis+for+Scalogram+analysis&btnG=&as_sdt=1%2C5&as_sdtp=#0. Accessed February 19, 2015.
22. van Schuur WH. Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory. *Polit Anal*. 2003;11(2):139-163. doi:10.1093/pan/mpg002.
23. Watkins MW. Determining Parallel Analysis Criteria. *J Mod Appl Stat Methods*. 2005;5(2):344-346. doi:10.22237/jmasm/1162354020.
24. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;2(4):433-459. doi:10.1002/wics.101.
25. Browne, M. W., & Cudeck R. Alternative ways of assessing model fit. 1993.
26. Gibbons C, Bower P, Lovell K, Valderas J, Skevington S, Bower P. Electronic Quality of Life Assessment Using Computer-Adaptive Testing. *J Med Internet Res*. 2016;18(9):e240. doi:10.2196/jmir.6053.
27. Hambleton R, Swaminathan H, Rogers H. *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage; 1991.
28. Wright B. Local dependency, correlations and principal components. *Rasch Meas Trans*. 1996;10(3):509-511.
29. Yen WM. Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Appl Psychol Meas*. 1984;8(2):125-145. doi:10.1177/014662168400800201.
30. Christiansen K, Maransky G, Horton M. Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Appl Psychol Meas*. 2017;41(3):179-194.
31. Lundgren N, medicine AT-J of rehabilitation, 2011 undefined. Past and present issues in Rasch analysis: the functional independence measure (FIM™) revisited. *europepmc.org*. http://europepmc.org/abstract/med/21947180. Accessed December 19, 2017.
32. Holland P, Wainer H. *Differential Item Functioning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.; 2012.
33. Teresi J. Different approaches to differential item functioning in health applications:

Advantages, disadvantages and some neglected topics. *Med Care*. 2006. http://journals.lww.com/lww-medicalcare/Abstract/2006/11001/Different_Approaches_to_Differential_Item.21.aspx. Accessed May 25, 2017.
34. Chalmers R. mirt: A multidimensional item response theory package for the R environment. *J Stat Softw*. 2012. https://scholar.google.co.uk/scholar?hl=en&q=Chalmers+MIRT&btnG=&as_sdt=1%2C5&as_sdtp=#0. Accessed April 11, 2016.
35. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care*. 2007;45(5):S22--S31. doi:10.1097/01.mlr.0000250483.85507.04.
36. Wainer H, Dorans N, Flaugher R, Green B. *Computerized Adaptive Testing: A Primer*.; 2000. https://books.google.co.uk/books?hl=en&lr=&id=73d9AwAAQBAJ&oi=fnd&pg=PP1&dq=Computerized+adaptive+testing:+A+Primer&ots=OtILaYiPRd&sig=y5BNptGGWjRqnwE3G9P4miUlzXo. Accessed October 31, 2016.
37. Choi SW, Swartz RJ. Comparison of CAT Item Selection Criteria for Polytomous Items. *Appl Psychol Meas*. 2009;33(6):419-440. doi:10.1177/0146621608327801.
38. Ark L Van der. Mokken scale analysis in R. *J Stat Softw*. 2007;20(11):1-19. http://www.jstatsoft.org/v20/a11/paper. Accessed February 13, 2015.
39. Choi S, Gibbons L, Crane P. Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo. *J Stat Softw*. 2011. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc3093114/. Accessed September 28, 2016.
40. RDC Team. R: A language and environment for statistical computing. http://www.r-project.org.
41. Revelle W. psych: Procedures for personality and psychological research. *Northwest Univ Evanston R Packag version*. 2014. https://scholar.google.co.uk/scholar?hl=en&q=psych%3A+Procedures+for+Personality+and+Psychological+Research&btnG=&as_sdt=1%2C5&as_sdtp=#0. Accessed February 13, 2015.
42. Carstensen B. The MethComp Package for R. In: *Comparing Clinical Measurement Methods*. Chichester, UK: John Wiley & Sons, Ltd; 2010:149-152. doi:10.1002/9780470683019.ch13.
43. Choi S. Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Appl Psychol Meas*. 2009. http://media.metrik.de/uploads/incoming/pub/Literatur/2009_Firestar-Computerized Adaptive Testing Simulation Program for Polytomous Item Response Theory Models, Choi.pdf. Accessed February 18, 2015.
44. Grundy AC, Bee P, Meade O, et al. Bringing meaning to user involvement in mental health care planning: a qualitative exploration of service user perspectives. *J Psychiatr Ment Health Nurs*. 2016;23(1):12-21. doi:10.1111/jpm.12275.
45. Cree L, Brooks HL, Berzins K, Fraser C, Lovell K, Bee P. Carers' experiences of involvement in care planning: a qualitative exploration of the facilitators and barriers to engagement with mental health services. *BMC Psychiatry*. 2015;15(1):208.

doi:10.1186/s12888-015-0590-y.
46. Millard C, Wessely S. Parity of esteem between mental and physical health. *BMJ*. 2014;349:g6821. doi:10.1136/BMJ.G6821.
47. Magis D, Raîche G. catR An R Package for Computerized Adaptive Testing. *Appl Psychol Meas*. 2011. http://apm.sagepub.com/content/35/7/576.short. Accessed September 29, 2016.
48. Psychometrics Centre. Concerto Adaptive Testing Platform. 2013.

Figure_1

Click here to download Figure Fig_1_final.png

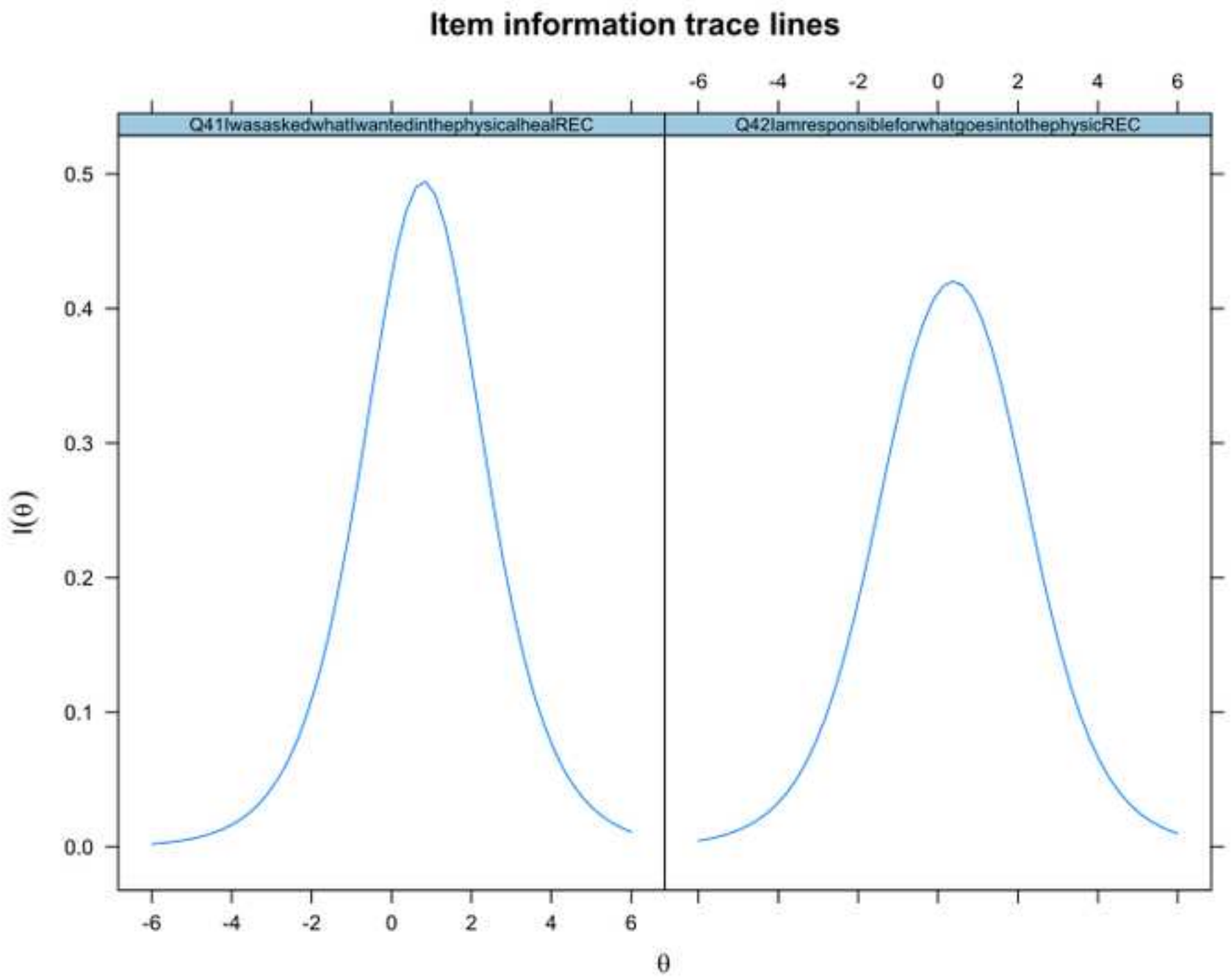


Figure_2

Click here to download Figure Figure_2_final.tiff

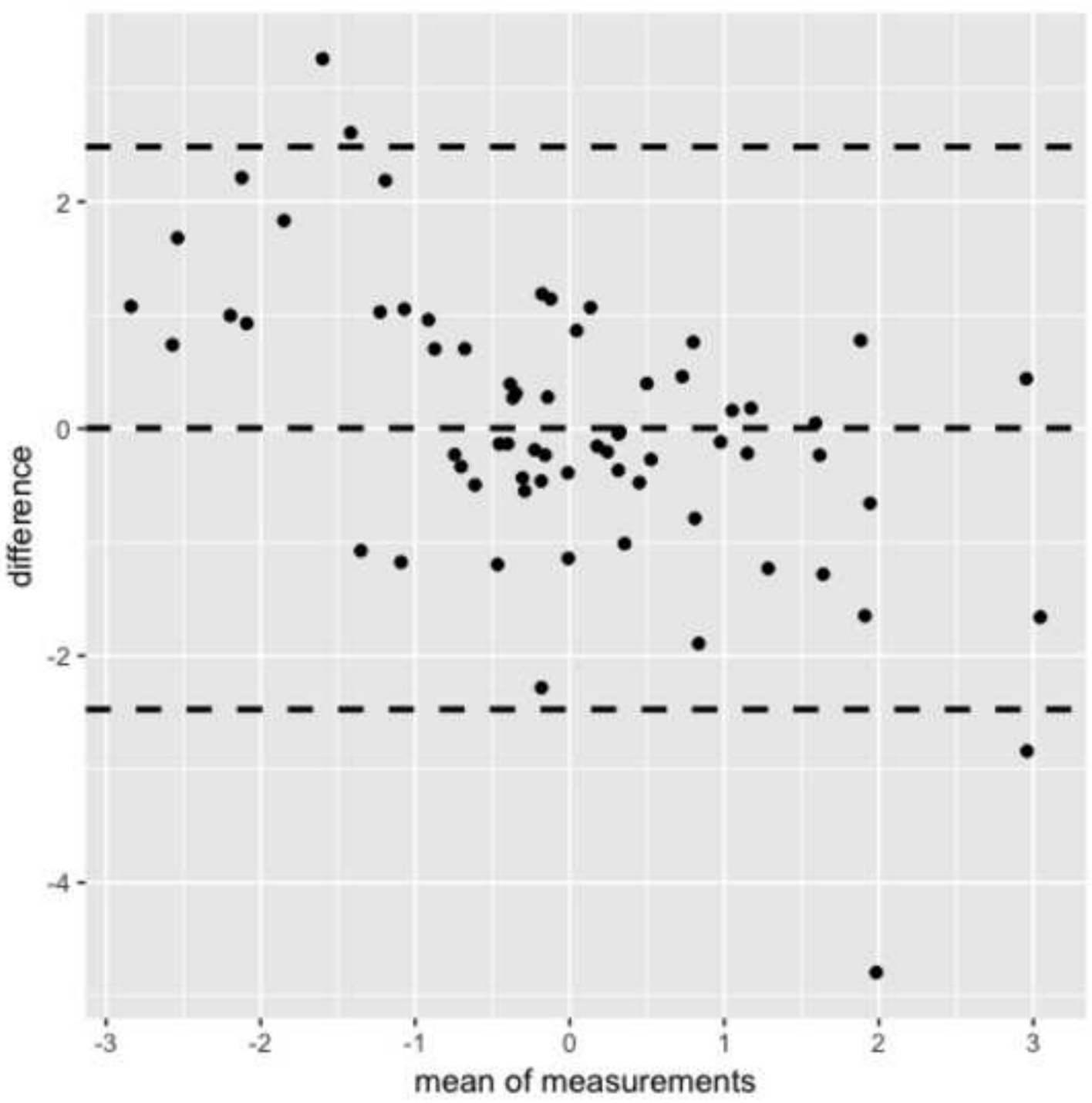
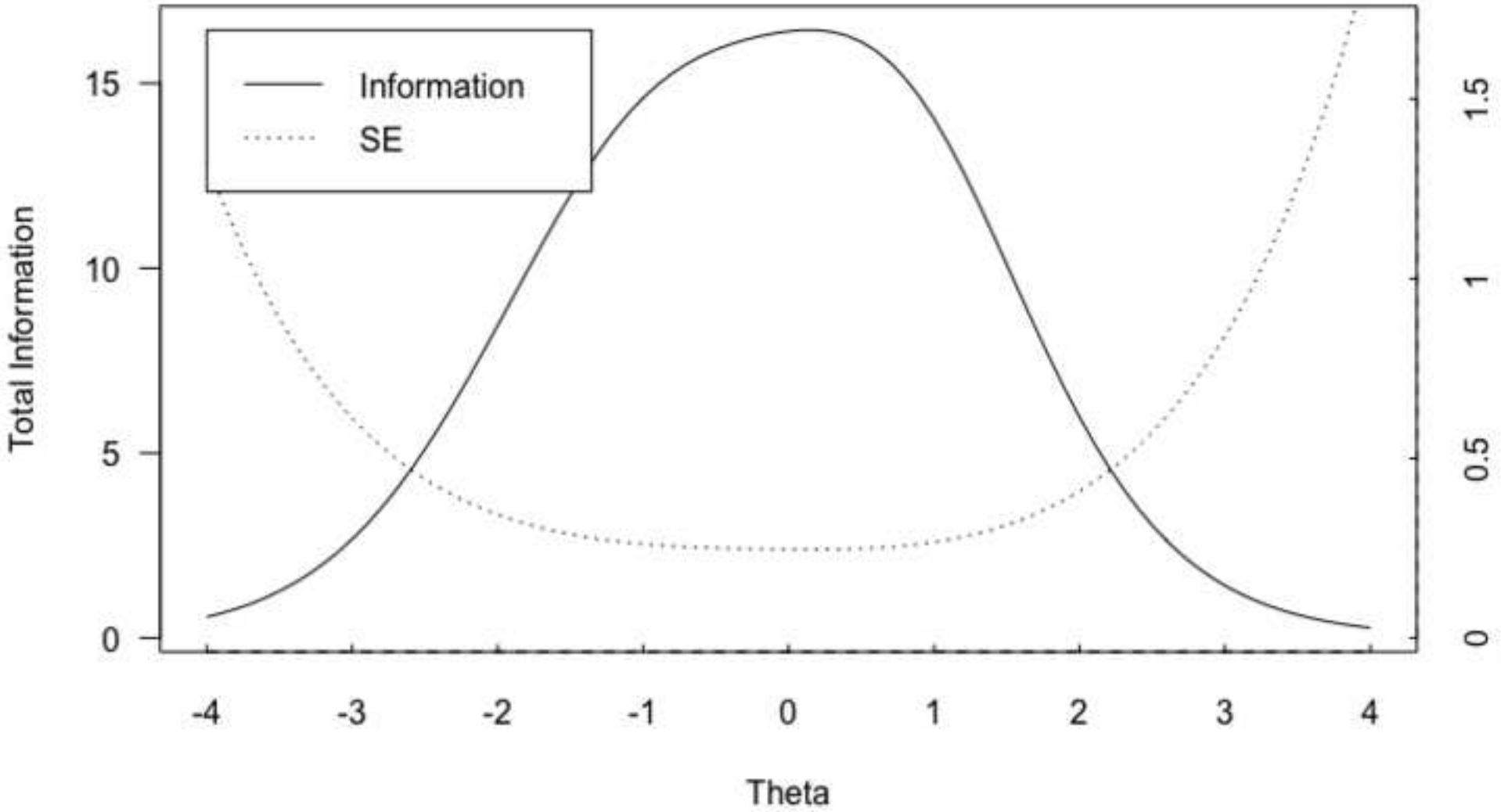


Figure_3

Click here to download Figure Fig_3_final.png



Click here to access/download
Supporting Information
retest_data.csv

Click here to access/download
Supporting Information
EQUIP_data.csv

ASSESSING MENTAL HEALTH SERVICE USER AND CARER INVOLVEMENT IN PHYSICAL HEALTH CARE PLANNING: THE DEVELOPMENT AND VALIDATION OF A NEW PATIENT-REPORTED ~~OUTCOME~~EXPERIENCE MEASURE.

Sidey-Gibbons CJ[1,2], Brooks H[3], Gellatly J[4], Small N[5], Lovell K[4], Bee P[4*]

1. Patient-reported Outcomes, Value, and Experience (PROVE) Center, Brigham and Women's Hospital, Boston, MA
2. Department of Surgery, Harvard Medical School, Boston MA.
3. Department of Psychological Sciences, Institute of Psychology, Health and Society, University of Liverpool, Liverpool, UK
4. Mental Health Research Group, Division of Nursing, Midwifery and Social Work, Faculty of Biology, Medicine and Health, School of Health Sciences, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK
5. NIHR School of Primary Care Research, Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK

*= Corresponding author.
e-mail = penny.bee@manchester.ac.uk

All authors contributed equally to this work.

Formatted: Font color: Auto
Formatted: Font color: Auto
Formatted: Font color: Auto
Formatted: Font color: Auto
Formatted: Font color: Auto

ABSTRACT

BACKGROUND

People living with serious mental health conditions experience increased morbidity due to physical health issues driven by medication side-effects and lifestyle factors. Coordinated mental and physical healthcare delivered in accordance with a care plan could help to reduce morbidity and mortality in this population. Efforts to develop new models of care are ~~held back~~hampered by a lack of validated instruments to accurately assess the extent to which mental health services users and ~~their~~ carers are involved in care planning for physical health.

OBJECTIVE

To develop a brief and accurate patient-reported ~~outcome~~experience measure (~~PROM~~PREM) capable of assessing involvement in physical health care planning for ~~users of~~ mental health ~~services~~ service users and their carers.

METHODS

We employed psychometric and statistical techniques to refine a bank of candidate questionnaire items, derived from qualitative interviews, into a valid and reliable measure

involvement in physical health care planning ~~for service users and carers.~~. We assessed the psychometric performance of the item bank using modern psychometric analyses. We assessed unidimensionality, scalability, fit to the partial credit Rasch model, category threshold ordering, local dependency, differential item functioning, and test-retest reliability. Once purified of poorly performing and erroneous items, we simulated computerized adaptive testing (CAT) with 15, 10 and 5 items using the calibrated item bank.

RESULTS

Issues with category threshold ordering, local dependency and differential item functioning were evident for a number of items in the nascent item bank and were resolved by removing problematic items. The final 19 item ~~PROM~~PREM had excellent fit to the Rasch model fit (x^2 = 192.94, df = 1515, *P*=.02, RMSEA = .03 (95% CI = .01-.04). The 19-item bank had excellent reliability (marginal r = 0.87). The correlation between questionnaire scores at baseline and 2-week follow-up was high (r = .70, *P* < .01) and 94.9% of assessment pairs were within the Bland Altman limits of agreement. Simulated CAT demonstrated that assessments could be made using as few as 10 items (mean SE = .43).

DISCUSSION

We developed a flexible patient reported outcome measure to quantify service user and carer involvement in physical health care planning. We demonstrate the potential to substantially reduce assessment length whilst maintaining reliability by utilizing CAT.

INTRODUCTION

People diagnosed with severe mental illnesses, such as disorder schizophrenia and ~~bi-polar~~bipolar disorders, exhibit higher rates of physical co-morbidities and, as a result, are significantly more likely to die prematurely than the general population.[1–3]

Factors contributing to this deterioration in physical health for mental health service users are known to include side-effects from anti-psychotic medications, higher rates of smoking and substance abuse, poor nutrition, and physical inactivity.[4] Though the relationship between serious mental health issues, physical cormorbidity, and reduced life expectancy is well understood, far less is known about how to organize care delivery ~~for this population~~ to improve physical health and reduce the risk of associated morbidity~~.~~ in this population. Recent ~~statistics~~evidence suggests that, despite ~~increasing~~increased awareness of these issues, mortality risk associated with all mental health conditions is rising internationally.[2]

One approach to improve the management of known risk factors is individualized care planning; [5,6] an approach which involves service users and carers working collaboratively with professionals to co-develop a written care plan ~~which~~. This plan aims to accurately ~~documents~~document the core issues that a service user would like to address as part of their mental health recovery. ~~The care plan is designed to be person-centred, actionable, and to make care providers accountable for providing care in line with the wishes of both service users and their carers.~~

A growing body of research ~~showsthat this sort of involvement is~~shows that, although collaborative care planning is s aligned with the desires of both service users and carers there is a paucity of care models which have been shown to effectively increase involvement in care planning for physical health in this way.[7] More broadly, increasing the quality of mental health services was the top research priority expressed by an international working group comprising both professionals as well as users and carers.[8]

Progress in the development of interventions to improve care planning involvement between service users, carers, and providers is stymied by the lack of a meaningful outcome assessment. Quantification of abstract subjective phenomena, such as involvement with care planning, is best accomplished by directly assessing the perspective of the service user or carer; usually using a tool commonly referred to as a patient-reported outcome measure (~~PROM~~PREM).

Patient reported outcome measures are a efficient and accurate way to quantify the views of service users and their carers. A relevant example is the EQUIP ~~PROM~~PREM, which was developed by our group to assess service user and carer involvement in mental health care planning.[9] Previous research has highlighted the importance of brief assessments for mental health service users and their carers, with a strong user preference for minimising response burden by developing shorter questionnaires.[9,10] New assessment modalities including computerized adaptive testing (CAT), are able to tailor person-centred assessments to the individual, a process which tends to result in shorter, more relevant assessments. [11]

The objective of the current paper is to create a novel ~~PROM~~PREM to assess mental health service user and carer involvement in physical health care planning. We seek to develop a ~~PROM~~PREM that is accurate, reliable, and suitable for individualized CAT assessment.

METHODS

ITEM DEVELOPMENT METHODS

A set of 67 candidate items were developed following qualitative interviews with mental health service users (SUs), their carers, and mental health professionals from the UK. Further details of the qualitative interview process can be found in a separate manuscript.[12] Items were developed to reflect six pre-identified themes; three of which covered general mental health care planning requirements and three of which were unique to physical health care planning. The general themes included: tailoring a collaborative working relationship between the service users and their carers and the service providers, maintaining a trusting relationship with a professional, having access to a tangible document which could be edited and updated. The physical health themes were: valuing physical health equally with mental health, experiencing coordinated care between health professionals in different disciplines, and having a personalised physical heath discussion.

DATA COLLECTION

Potential participants who expressed an interest in taking part in the study were given a participant information sheet written to current UK National Research Ethics Service (NRES)

guidelines. We worked with service users and carers to co-develop the information sheet. The information sheet included details on the study including the potential risks and benefits of taking part, the ways in which participants could take part in the study (e.g. online via SelectSurvey or through the completion and return of paper versions of the questionnaire), and provided potential participants with the contact details of researchers should they wish to discuss their involvement prior to taking part. All participants responded affirmatively to the question "I have read and understood the participant information sheet" and consent was implied by the completion and return of questionnaires. ~~Consent was implied by the completion and return of questionnaires.~~ The study and all associated procedures were approved by the London – West London and Gene Therapy Advisory Committee (GTAC) Research Ethics Committee (16/LO/0386) in February 2016.

Data analysis

We fitted data from nascent scale to the partial credit "Rasch" model (PCM)[13,14] in order to assess psychometric performance. We evaluated factor structure, scalability and monotonicity by fitting data to non-parametric Mokken model before more rigorous psychometric assessments using the PCM.[15] The combination of the two methodologies has been shown to be useful in previous research conducted by members of our group and others.[16–19]

Where scale data did not conform to the assumptions of either the Mokken or the partial credit model, an iterative process of item reduction was undertaken to remove the violating items from the analysis.[20]

The iterative process involved assessments of scalability, model and item fit to the PCM, category threshold disordering, local dependency, and differential item functioning (DIF). Each concept and the method by which it is assessed is described in greater detail below.

Mokken Analysis

The Mokken model is a non-parametric extension of the simple deterministic Guttman scaling model. [21] The model provides a framework to extend the unreastically error-free Guttman models using probabilistic estimation, thus accounting for measurement error.[22] As a non-parametric item response theory (NIRT) model, the Mokken models relax some assumptions of item response theory whilst affirming essential assumptions such as unidimensionality and scalability.[22] We fitted data to the double monotonicity model, a NIRT model which estimates a single parameter for each item (*i.e.,* the level of the construct which that item assesses). By successfully fitting scale data to a Mokken model it can be said to be both unidimensional and properly scaled. We utilized parallel polychoric principal component analysis which compared the experimental eigenvalues with a Monte Carlo simulated eigenvalues to verify the unidimensional factor structure before proceeding to item response theory analysis.[23,24]

The Partial Credit Model

The PCM is a measurement model which describes the probabalistic relationship between the assessment and the respondent as an interaction between the amount of the latent construct that the respondent has (*i.e.* involvement with physical health care planning) and the level of the latent construct which the item measures. Both the amount of the construct that the respondent has and the level of the latent construct that the item measures can be described in terms of theta (θ). For example, a item which measures a very high level of

physical health care planning (which would be an question that we would not expect many people to affirm; for example the questionnaire item may ask about service user or carer's access to ~~an~~a document containing a detailed strategy for physical health care) would be less likely to be affirmed than an item measuring a low level of the latent construct (which would be a question that we would expect many people to affirm; for example the questionnaire item may ask about whether a health care professional had asked whether a service user was receiving any care for physical health issues).

Goodness-of-fit statistics can be used to assess the data's fit to the PCM model at both the item and scale level. In this study we used both the Chi-square and root-mean square error of approximation ~~(RMSEA > .05)~~ to evaluate the fit to the model. We accepted both a non-significant Chi-square interaction (P > .05) and RMSEA (< .05) indicating good fit. [25]

Category Threshold Ordering

In the case of a Likert or 'multiple choice' item response the probability of responding to each category is modelled separately. As the level of the underlying construct (*i.e.,* involvement in physical health care planning) rises the probability of responding to each Likert category rises to a peak before falling. Different probabilities are given for each response category at every level of θ. It is essential that each category becomes the most likely response option at a certain level of θ. If this is not the case the item is said to exhibit category threshold disordering.

Category threshold disordering refers to the situation in which one or more of the Likert response categories are not the most likely response at any point of along the underlying θ continuum. In the case of disordered category thresholds, we 'collapsed' adjacent categories so that they received the same score. Care was taken not to collapse categories if it were semantically illogical to do so, (*i.e.*, "Agree" would not be collapsed into "Neither Agree nor Disagree"). An illustrated side-by-side example is provided in a previous paper from our group. [26]

Item response theory models (of which the Rasch model is a special case) are predicated on the assumption that differences in responses to items are driven solely by changes in the underlying trait.[27] One way in which items can violate this assumption is local dependency, a situation whereby the response to one item is dependent on the response to another.[28] In practice, this can occur where items are too similar. Local dependency is assessed using Yen's Q3 statistic, in which the correlation of item residuals are compared, and item pairs with residual correlations beyond a threshold are said to be locally dependent. We set the threshold to be equal to .2 + the average observed residual.[29,30]

Local dependency can be resolved in a number of ways, including subtesting (where locally dependent items are joined into a 'super' item) and item deletion.[31] As we began with a large bank of candidate items, we elected to remove items which were locally dependant. Our strategy was to remove an item if it were locally dependent with more than one other item and then, in the case that a locally independent item pair only demonstrated dependency with one another, item information curves for each item were examined alongside the item wording and the item which provided less information was removed from further analysis.

Another issue which can interfere with the assumption that differences in item scores ought to be driven solely by differences in the underlying trait is differential item functioning (DIF).[32] Differential item functioning occurs when the probability of a certain response to a question varies across different demographic groups. For example, if men were more likely to respond affirmatively to a certain item than women despite having an equal level of overall involvment with physical health care planning, that question would be said to be affected by DIF. We used the iterative hybrid ordinal logistic regression/item response theory approach to conduct DIF analyses. For items flagged as having signficiant DIF following Bonferroni correction, we used the McFadden pseudo R2 estimation with recommend cut-off of R2 > .035 being indicative of meaningful DIF.[33] By assessing DIF between service users and carers we will explore the suitablility of the nascent ~~PROM~~PREM for both groups. Models were fitted with missing data present. However, missing data were imputed using IRT-based estimation.[34] Given the well-documented issues with model fit statistics, we prioritized meeting the assumptions of the Mokken and Rasch models over model fit, as has been recommended elsewhere.[35]

Items that violated any of the above assumptions were removed, and the remaining items were reanalyzed. We evaluated the reliability of the final scale and the overall fit to the Rasch model. Once a final set of purified questions were calibrated by fitting them to the PCM, we simulated computerized adaptive tests (CATs)[36]. The CAT algorithms conducted stepwise assessments by iteratively selecting the item which will maximise the test information based on the participant based on participant's θ estimate which is based off their previous responses. The first item for the assessment is the item which maximises information at the mean population level of θ.

We simulated CATs to assess the viability of brief assessment using the nascent scale using the final items as a 'bank' of candidate items. In computerized adaptive testing an algorithm is used to select the next most appropriate item for the patients based on their previous responses. This approach has been shown to substantially reduce the length of tick-box assessments whilst maintaining, and even increasing, reliability.[18,26]

We similated CATs using the Firestar script for R. Firestar uses a Bayesian expected a posteriori θ estimator and selected items based on the maximum posterior weighted information (MPWI) criterion. The MPWI selects items based on the item information weighted by the posterior distribution of trait/phenomena values.[37]This criterion has been shown to provide excellent measurement information for CAT using polytomous items.

Software

Analyses were conducted using the R Statistical Computing Environment with the 'mokken', 'mirt', 'lordif', 'psych', 'ggplot2', 'methcomp' and 'BlandAltLeh' packages installed. [34,38–42]. Computer adaptive testing simulation was conducted using the FIRESTAR script, which was modified to add additional statistics.[43] This modified FIRESTAR code is available on request from the authors.

Results

We collected data from 267 mental health services users from the United Kingdom. 67 participants completed the 67 candidate questionnaire items a second time after two weeks. No data were available on the number of participants who began the survey but did not complete it. ~~Missing responses were given in 2966 or 18291 cells (16%).~~16% of PREM data was missing.

Table 1 Demographic information for 267 participants recruited into the study

Age		44(14)
Gender	~~206~~69.3% Female	~~62~~ ~~Male~~
		21% Male
		9.7% unreported
Ethnicity		
	83.7% White	~~83.70%~~
		16.3% Non-white
Service user/carer status		
SU		~~196~~66%
Carer		~~46~~15%
~~SUandC~~Both SU and carer		~~33~~12%
Not reported		7%
Geographic location		
Northern England		32%
The Midlands		23.7%
Southern England		39.4%
Ireland		1%
Scotland		1.5%
Wales		2.5%

Key: SU = Service users~~, C = Carer~~

MOKKEN ANALYSIS

Mokken analysis revealed violations of monotonicity for a number of items (5, 6, 8, 10, 25, 26, 40). In addition, Loevinger's scalability coefficient was too low (Item H >.30) for items 7, 9, 39. The 57 remaining items were free from violations of monotonicity and were unidimensional. Parallel principal component and factor analysis confirmed the unidimensional structure of the dataset as the eigenvalue for the second factor/component (2.87, 2.16) was below simulated eigenvalues in the Monte Carlo dataset (1.50, 1.19).

Deleted Cells
Formatted: Font: Calibri, 11 pt, Font color: Dark Blue
Formatted: Left
Formatted Table
Formatted: Right
Formatted: Right
Formatted: Font: Calibri, 11 pt, Font color: Dark Blue
Formatted: Font: Calibri, 11 pt, Font color: Dark Blue
Formatted: Left
Deleted Cells
Deleted Cells
Inserted Cells
Formatted Table
Formatted: Right
Formatted: Font: Calibri, 11 pt, Font color: Dark Blue
Deleted Cells
Formatted: Right
Formatted: Font: Calibri, 11 pt, Font color: Dark Blue
Formatted: Font: Calibri, 11 pt, Font color: Dark Blue
Formatted: Left
Formatted Table
Formatted: Right
Formatted: Font: Calibri, 11 pt, Font color: Dark Blue
Formatted: Font: Calibri, 11 pt, Font color: Dark Blue
Formatted: Left
Formatted: Left
Formatted: Font: Calibri, 11 pt, Font color: Dark Blue
Formatted: Right
Formatted: Font: Calibri, 11 pt, Font color: Dark Blue

RASCH ANALYSIS

The remaining items were fitted to the partial credit model. The initial fit to the model was poor (RMSEA > .10); thus prompting evaluation of item performance in the context of Rasch model assumptions. There appeared to be substantial threshold disordering throughout the scale. A single solution was chosen to rescore all items (0-1-1-2-2). The amended threshold probability curves for all the items can be see in Appendix 1.

Model fit improved slightly after rescoring but was still unacceptable (RMSEA = .097 (95% CI = .091-.10)). We evaluated the correlations between item residuals, which were above the threshold in a number of instances. In total, 96 item pairs were locally dependant (see Appendix 2). A total of 27 individual items that displayed local dependency with more than one other item ~~and~~ were removed from the analysis. Four sets of items remained which were locally dependant with one another. The item information curves for both pairs of locally-dependent items were compared side-by-side (see Fig 2) and in each case the item with the lowest item information removed from the scale.

Figure 1. Comparison of item information curves for locally dependent items

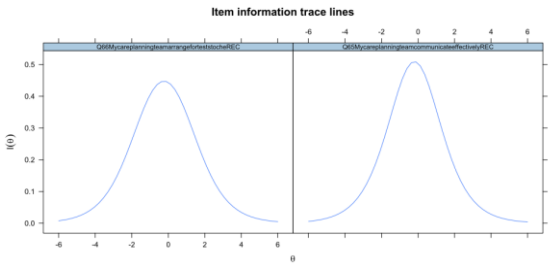


Figure 2. Shows item information curves for a pair of locally-dependent items. The amount of information which each gives about the participant is given on the *y*-axis and the level of underlying construct that the person has (i.e., involvement with physical health care planning) is on the *x*-axis.

Formatted: Font: 12 pt

No DIF was detected for age or gender but item 65 “My care planning team communicates effectively” was found to have significant DIF (R2 change = .06) between service users and carers.

Following adjustment for category threshold ordering, local dependency, and differential item functioning; item 36 “I feel comfortable attending discussions about my care plan” misfit the model and was removed (x^2 = 34.81, df = 15, *P*=.003). The removal of item 36 led to a final item bank of 19 items which were free from breaches of assumptions of the Rasch model, displayed excellent model fit (x^2 = 192.94, df = 1515, *P*=.02, RMSEA = .03 (95% CI = .01-.04). The 19-item bank had excellent reliability (marginal r = 0.87).

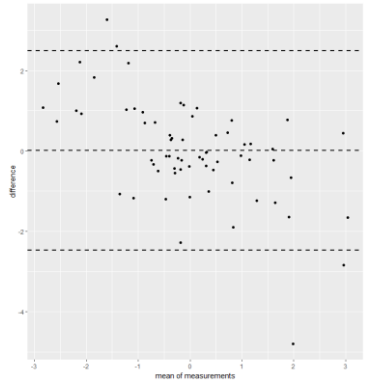
Table 2 Details of final items and item threshold parameters

Item Number	Original number	Wording	Model fit			Item Threshold Parameters		item fit statistics			
			χ^2	df	P	delta 1	delta 2	χ^2	df	P	Scoring
1	50	My care planning team ask about my existing physical health conditions.	21.81	15	.11	-1.20	.12	30.8	34	0.63	0-1-1-2-2
2	24	The physical health information in my care plan is personalised.	13.29	17	.72	-1.92	.20	38.07	34	0.29	0-1-1-2-2
3	53	My care planning team encourage me to take responsibility for my physical health care planning.	14.10	16	.59	-2.01	.27	24.53	32	0.82	0-1-1-2-2
4	37	My opinion on my physical health is valued by my care planning team.	13.48	16	.64	-2.03	.38	26.98	31	0.67	0-1-1-2-2
5	4	I know who reads the physical health information contained within my care plan	11.79	15	.69	-1.02	.43	27.49	33	0.74	0-1-1-2-2
6	55	My care planning team offer practical advice about my physical health.	15.44	15	.42	-1.42	.51	33.18	31	0.36	0-1-1-2-2
7	13	My care plan gives details of my physical health history.	20.20	15	.16	-1.16	.52	26.99	31	0.67	0-1-1-2-2
8	15	My thoughts about my physical health are included in my care plan.	11.20	17	.85	-.62	.76	28.12	29	0.51	0-1-1-2-2
9	52	I experience continuity of care for the treatment of both my physical health conditions and mental health conditions.	15.22	17	.58	-.68	.77	39.73	32	0.16	0-1-1-2-2
10	22	The physical health information in my care plan is helpful.	9.43	19	.97	-1.61	.83	28.01	32	0.67	0-1-1-2-2
11	16	Physical health reviews are carried out in a timely manner.	14.55	16	.56	-.53	.86	28.02	32	0.67	0-1-1-2-2
12	62	My care planning team have a good understanding of my fears about future physical health conditions.	15.30	18	.64	-.77	.92	28.4	32	0.65	0-1-1-2-2
13	56	My care planning team have the time they need to talk to me about physical health concerns.	28.39	19	.08	-1.18	.94	27.15	28	0.51	0-1-1-2-2
14	44	The content of my physical health care plan is responsive to changes in my circumstances.	15.06	19	.72	-1.14	1.06	26.01	33	0.8	0-1-1-2-2
15	27	Information in my care plan has helped me to maintain my physical health.	13.12	18	.78	-.63	1.08	26.09	33	0.8	0-1-1-2-2
16	41	I was asked what I wanted in the physical health information in my care plan.	14.07	16	.59	-.25	1.13	32.37	34	0.55	0-1-1-2-2
17	60	The care plan adequately addresses any side effects I experience from my medication.	14.79	19	.74	-1.22	1.13	40.21	35	0.25	0-1-1-2-2
18	46	I have had the opportunity to invite all the relevant people to care planning meetings related to my physical health.	17.23	16	.37	-.55	1.27	31.57	33	0.54	0-1-1-2-2
19	51	My care planning team makes sure my mental health is not prioritised over my physical health.	21.01	19	.34	-1.15	1.51	31.72	32	0.48	0-1-1-2-2

TEST-RETEST RELIABILITY

The correlation between theta scores at baseline and 2-week follow-up was high (r = .70, $P <$ .01). Bland Altman analysis revealed that 94.9% of assessment pairs were within the 95% limits of agreement (see Figure 2).

Figure 2 – Bland Altman plot for test-retest reliability



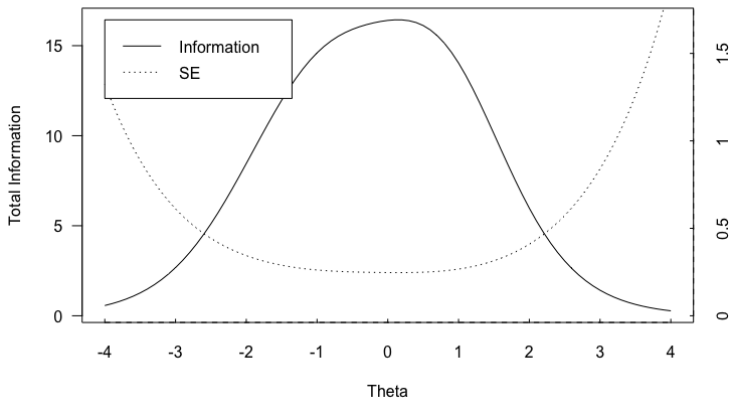
COMPUTERIZED ADAPTIVE TESTING

Adaptive testing simulations were conducted with a simulated Gaussian N (-0.08,1.90) distribution, which matched the ~~sample~~ distribution of the data used to develop the item banks. Results of CAT simulation are shown in the Table 3. Assessments as short as 10 items demonstrated high correlation with the total score of the full scale. The overall information and standard error which was available in the entire item banks is displayed in Figure 3.

Table 3. Summary of simulated computer adaptive tests.

	Standard Error (SE)			
Number of items	Mean	SD	Range	Correlation with full scale
19	.36	.04	.32-.56	1
15	.41	.04	.38-.60	.98
10	.43	.04	.40-.65	.95
5	.66	.04	.62-.78	.87

Figure 3 – Overall Scale Information and Standard Error
Key – SE = Standard Error



DISCUSSION

We present the co-development and validation of a new service user and carer-reported assessment for physical health care planning in serious mental health services, the EQUIP Physical Health ~~PROM~~PREM (EQUIP-PH-~~PROM~~PREM).

The new ~~PROM~~PREM contains 19 items which were successfully fitted to a single-parameter Rasch item response theory model. The ~~PROM~~PREM is suitable for assessing both service users and carers. The 19 items also serve as an item bank for computerized adaptive testing (CAT) which can tailor assessments to the individuals who complete the ~~PROM~~PREM. We show that using CAT administration could substantially reduce burden of response by reducing the number of items in the assessment from 19 to 10, whilst still maintaining acceptable accuracy and high correlation between scores.

In the EQUIP-PH-~~PROM~~PREM we provide a tool to support investigations into the experience of service users and carers who are receiving care for a ~~serious~~severe mental illness from mental health providers. Adequate service user and carer involvement in care planning decisions are predicated on successful interaction both within and between stakeholder groups. In order to ensure the new ~~PROM~~PREM incorporated these important aspects the items were developed in collaboration with service users, carers and mental health professionals.

The final ~~PROM~~PREM items include those that cover having the opportunity and time to be able to discuss physical health concerns, reflecting previously identified organisational barriers to providing integrated care.[12] Similarly, they highlight the importance of *co-created* care plans, which are known to be highly valued by both service users and carers [44,45] Further items serve to facilitate long-term self-management skills that are required to manage physical health concerns.

This new ~~PROM~~PREM has operationalized the evidence-based best practice framework developed previously which will allow health care providers and service users to challenge current practice by quantifying service user and carer involvement from the user perspective.[12]

The measure will facilitate benchmarking of service quality and service user experience, aligned with contemporary philosophies and policies for collaborative recovery-focused mental health care. The philosophy of the new ~~PROM~~PREM is that mental and physical health are equally important (the so-called parity of esteem), and parity of esteem is increasingly being embedded in policy and practice imperatives derived from stakeholder consultation.[46]

The EQUIP-PH ~~PROM~~PREM assesses issues which have been consistently highlighted in consultations with service users and carers and, as such, is well suited for use as an ~~assessment tool for interventions to provide all health care personnel with an understanding of both physical and mental health during professional training to ensure that one type of health need is not prioritised over the other.~~tool to assess the outcome of interventions. Other relevant interventions include those designed to improve inter- and intra- professional communication including professional training and improved health systems to enhance the integration and continuity of care for those under the care of health services.

The current study has ~~a few~~some limitations. Firstly, our dataset consisted of predominantly white, female service users. Though all systematic differences between demographic groups were corrected for in the current analysis, further research would be warranted to ensure that the items perform well in groups which were not well represented in our ~~sample~~data. It should be noted that whilst we demonstrated uniform scale performance across demographic groups – including ~~SUs~~service users and carers, we did not collect information relating to comorbidities, physical activity or substance and further research would be necessary to explicitly confirm that the scale is unaffected by differences in disease or lifestyle factors within groups of ~~SUs.~~service users.

Our study is also limited by the necessity to evaluate the CATs using simulated, rather than actual, data. This technique is likely to the slightly over-estimate the accuracy of the CAT as it does not take into account aberrant responders who do not conform to the expectations of the model. Our previous research developing item banks for depression and quality of life suggests that this effect is marginal and that CAT assessment is efficient and precise both when simulations are made using participant data and when the CAT is deployed in the real world.[18]

It is noteworthy that when administering CATs each individual respondent is likely to complete different combinations of items which form a subset of the complete item bank. Though the scores between the unidimensional CAT and the fixed-length short-form are highly correlated, there is no guarantee that every patient will complete items from each of the content domains which were nominated by service users and carers. In the current manuscript, we prioritize brevity and accuracy and simulate CAT administration without

Formatted: Font color: Text 1

Formatted: Font color: Text 1

content balancing or prioritizing certain items. We acknowledge that other users may prioritize item exposure and thus may utilize CATs differently.

Parties who wish to use CAT administration for the EQUIP-PH measure are directed towards many ~~excellent~~ packages available for the R Statistical Programming Environment including mirt and catR.[34,47] ~~We believe the most straightforward tool for developing and deploying assessments which include CAT~~One tool for implementing CATs is the Concerto platform, developed and maintained by the University of Cambridge.[~~48~~48] Further details can be found on the Concerto website (concertoplatform.com) or by request to the authors of this manuscript.

Our study also has some notable strengths. We have collected a geographically diverse ~~sample~~group of both ~~SUs~~service users and carers and created a flexible assessment which can be used without modification of assessing and comparing both groups. The EQUIP-PH ~~PROM~~PREM which we have developed is related to the EQUIP ~~PROM~~measure, a questionnaire measure for service user and carer involvement in care planning, which was recently developed by our group[9]. Both tools could be used together to gain a holistic understanding of how involved service users and carers are in ~~global~~mental health care planning. Further research could usefully be conducted to understand the scores from the two instruments in relation to one another and provide further insight into their use as a tool to assess global care planning and service delivery.

In conclusion, The EQUIP-PH ~~PROM~~PREM is a brief, accurate, and flexible service user- and carer-reported assessment ~~for~~of involvement in physical health care planning for users of mental health services with serious mental ~~health services~~illnesses. The measure provides a reliable means to evaluate and benchmark the quality of physical health management in the context of mental health care.

Acknowledgement

This research was funded by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care (NIHR CLAHRC) Greater Manchester. The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supporting information.

Supporting Information 1 : Anonymized baseline dataset
Supporting Information 2 : Anonymized follow up dataset

References

1. Harris EC, Barraclough B. Excess mortality of mental disorder. *Br J Psychiatry*. 1998;173(1):11-53. doi:10.1192/BJP.173.1.11.
2. Walker ER, McGee RE, Druss BG. Mortality in Mental Disorders and Global Disease Burden Implications. *JAMA Psychiatry*. 2015;72(4):334. doi:10.1001/jamapsychiatry.2014.2502.
3. Rodgers M, Dalton J, Harden M, Street A, Parker G. Integrated care to address the

Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Text 1
Formatted: Font color: Auto

physical health needs of people with severe mental illness: a rapid review. 2016. https://www.ncbi.nlm.nih.gov/books/NBK355962/. Accessed December 18, 2017.
4. Brown S, Birtwistle J, Roe L, Thompson C. The unhealthy lifestyle of people with schizophrenia. *Psychol Med*. 1999;29(3):697-701. doi:10.1017/S0033291798008186.
5. Doyle C, Lennox L, open DB-B, 2013 undefined. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *bmjopen.bmj.com*. http://bmjopen.bmj.com/content/3/1/e001570.short. Accessed December 18, 2017.
6. Care Quality Commission. *Survey of Mental Health Inpatient Services*.; 2009.
7. Bee P, Brooks H, Fraser C, Lovell K. Professional perspectives on service user and carer involvement in mental health care planning: a qualitative study. *Int J Nurs Stud*. 2015;52(12):1834-1835. http://www.sciencedirect.com/science/article/pii/S0020748915002308. Accessed December 18, 2017.
8. Fiorillo A, Luciano M, Del Vecchio V, Sampogna G, Obradors-Tarragó C, Maj M. Priorities for mental health research in Europe: A survey among national stakeholders' associations within the ROAMER project. *World Psychiatry*. 2013;12(2):165-170. doi:10.1002/wps.20052.
9. Bee P, Gibbons C, Callaghan P, Fraser C, Lovell K. Evaluating and Quantifying User and Carer Involvement in Mental Health Care Planning (EQUIP): Co-Development of a New Patient-Reported Outcome Measure. *PLoS One*. 2016;11(3):e0149973. doi:10.1371/journal.pone.0149973.
10. Gibbons CJ, Bee PE, Walker L, Price O, Lovell K. Service user- and carer-reported measures of involvement in mental health care planning: methodological quality and acceptability to users. *Front psychiatry*. 2014;5:178. doi:10.3389/fpsyt.2014.00178.
11. Gibbons C, Bower P, Lovell K, Valderas J, Skevington S. Electronic quality of life assessment using computer-adaptive testing. *J Med Internet Res*. 2016;18(9):e240.
12. Small N, Brooks H, Grundy A, et al. Understanding experiences of and preferences for service user and carer involvement in physical health care discussions within mental health care planning. *BMC Psychiatry*. 2017;17(1):138. doi:10.1186/s12888-017-1287-1.
13. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educational Research; 1960.
14. Masters GN. A rasch model for partial credit scoring. *Psychometrika*. 1982;47(2):149-174. doi:10.1007/BF02296272.
15. Mokken RJ. *A Theory and Procedure of Scale Analysis: With Applications in Political Research*. Walter de Gruyter; 1971. https://books.google.com/books?hl=en&lr=&id=vAumIrkzYj8C&pgis=1. Accessed February 16, 2015.
16. Meijer RR, Sijtsma K, Smid NG. Theoretical and Empirical Comparison of the Mokken and the Rasch Approach to IRT. *Appl Psychol Meas*. 1990;14(3):283-298. doi:10.1177/014662169001400306.
17. Stochl J, Jones PB, Croudace TJ. Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC Med Res Methodol*. 2012;12(1):74. doi:10.1186/1471-2288-12-74.
18. Loe BS, Stillwell D, Gibbons C. Computerized Adaptive Testing Provides Reliable and

Efficient Depression Measurement Using the CES-D Scale. *J Med Internet Res*. 2017;19(9):e302. doi:10.2196/jmir.7453.
19. Gibbons CJ, Small N, Rick J, Burt J, Hann M, Bower P. The Patient Assessment of Chronic Illness Care produces measurements along a single dimension: results from a Mokken analysis. *Health Qual Life Outcomes*. 2017;15(1):61. doi:10.1186/s12955-017-0638-4.
20. Pallant J, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol*. 2007;46(1):1-18.
21. Guttman L. The basis for scalogram analysis. 1949. https://scholar.google.co.uk/scholar?hl=en&q=The+basis+for+Scalogram+analysis&btnG=&as_sdt=1%2C5&as_sdtp=#0. Accessed February 19, 2015.
22. van Schuur WH. Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory. *Polit Anal*. 2003;11(2):139-163. doi:10.1093/pan/mpg002.
23. Watkins MW. Determining Parallel Analysis Criteria. *J Mod Appl Stat Methods*. 2005;5(2):344-346. doi:10.22237/jmasm/1162354020.
24. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscip Rev Comput Stat*. 2010;2(4):433-459. doi:10.1002/wics.101.
25. Browne, M. W., & Cudeck R. Alternative ways of assessing model fit. 1993.
26. Gibbons C, Bower P, Lovell K, Valderas J, Skevington S, Bower P. Electronic Quality of Life Assessment Using Computer-Adaptive Testing. *J Med Internet Res*. 2016;18(9):e240. doi:10.2196/jmir.6053.
27. Hambleton R, Swaminathan H, Rogers H. *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage; 1991.
28. Wright B. Local dependency, correlations and principal components. *Rasch Meas Trans*. 1996;10(3):509-511.
29. Yen WM. Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Appl Psychol Meas*. 1984;8(2):125-145. doi:10.1177/014662168400800201.
30. Christiansen K, Maransky G, Horton M. Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch Model Using Residual Correlations. *Appl Psychol Meas*. 2017;41(3):179-194.
31. Lundgren N, medicine AT-J of rehabilitation, 2011 undefined. Past and present issues in Rasch analysis: the functional independence measure (FIM™) revisited. *europepmc.org*. http://europepmc.org/abstract/med/21947180. Accessed December 19, 2017.
32. Holland P, Wainer H. *Differential Item Functioning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.; 2012.
33. Teresi J. Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Med Care*. 2006. http://journals.lww.com/lww-medicalcare/Abstract/2006/11001/Different_Approaches_to_Differential_Item.21.aspx. Accessed May 25, 2017.
34. Chalmers R. mirt: A multidimensional item response theory package for the R environment. *J Stat Softw*. 2012. https://scholar.google.co.uk/scholar?hl=en&q=Chalmers+MIRT&btnG=&as_sdt=1%2C5&as_sdtp=#0. Accessed April 11, 2016.

35. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Med Care*. 2007;45(5):S22--S31. doi:10.1097/01.mlr.0000250483.85507.04.
36. Wainer H, Dorans N, Flaugher R, Green B. *Computerized Adaptive Testing: A Primer*.; 2000. https://books.google.co.uk/books?hl=en&lr=&id=73d9AwAAQBAJ&oi=fnd&pg=PP1&dq=Computerized+adaptive+testing:+A+Primer&ots=OtILaYiPRd&sig=y5BNptGGWjRqnwE3G9P4miUlzXo. Accessed October 31, 2016.
37. Choi SW, Swartz RJ. Comparison of CAT Item Selection Criteria for Polytomous Items. *Appl Psychol Meas*. 2009;33(6):419-440. doi:10.1177/0146621608327801.
38. Ark L Van der. Mokken scale analysis in R. *J Stat Softw*. 2007;20(11):1-19. http://www.jstatsoft.org/v20/a11/paper. Accessed February 13, 2015.
39. Choi S, Gibbons L, Crane P. Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo. *J Stat Softw*. 2011. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc3093114/. Accessed September 28, 2016.
40. RDC Team. R: A language and environment for statistical computing. http://www.r-project.org.
41. Revelle W. psych: Procedures for personality and psychological research. *Northwest Univ Evanston R Packag version*. 2014. https://scholar.google.co.uk/scholar?hl=en&q=psych%3A+Procedures+for+Personality+and+Psychological+Research&btnG=&as_sdt=1%2C5&as_sdtp=#0. Accessed February 13, 2015.
42. Carstensen B. The MethComp Package for R. In: *Comparing Clinical Measurement Methods*. Chichester, UK: John Wiley & Sons, Ltd; 2010:149-152. doi:10.1002/9780470683019.ch13.
43. Choi S. Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Appl Psychol Meas*. 2009. http://media.metrik.de/uploads/incoming/pub/Literatur/2009_Firestar-Computerized Adaptive Testing Simulation Program for Polytomous Item Response Theory Models, Choi.pdf. Accessed February 18, 2015.
44. Grundy AC, Bee P, Meade O, et al. Bringing meaning to user involvement in mental health care planning: a qualitative exploration of service user perspectives. *J Psychiatr Ment Health Nurs*. 2016;23(1):12-21. doi:10.1111/jpm.12275.
45. Cree L, Brooks HL, Berzins K, Fraser C, Lovell K, Bee P. Carers' experiences of involvement in care planning: a qualitative exploration of the facilitators and barriers to engagement with mental health services. *BMC Psychiatry*. 2015;15(1):208. doi:10.1186/s12888-015-0590-y.
46. Millard C, Wessely S. Parity of esteem between mental and physical health. *BMJ*. 2014;349:g6821. doi:10.1136/BMJ.G6821.
47. Magis D, Raîche G. catR An R Package for Computerized Adaptive Testing. *Appl Psychol Meas*. 2011. http://apm.sagepub.com/content/35/7/576.short. Accessed September 29, 2016.
48. Psychometrics Centre. Concerto Adaptive Testing Platform. 2013.