Development and validation of the Symptom Burden Questionnaire™ for Long Covid: A Rasch analysis

Sarah E Hughes1,2,3,4*, Shamil Haroon4, Anuradha Subramanian4, Christel McMullan1,5,7, Olalekan L Aiyegbusi1,2,3,4, Grace M Turner1,4, Louise Jackson4, Elin Haf Davies6, Chris Frost8, Gary McNamara8, Gary Price1, Karen Matthews9,10, Jennifer Camaradou10,11, Jane Ormerod12, Anita Walker1, Melanie J Calvert1,6,15-18

1Centre for Patient Reported Outcomes Research, Institute of Applied Health Research, University of Birmingham, Birmingham, UK
2National Institute for Health Research (NIHR) Applied Research Collaboration (ARC) – West Midlands, Birmingham, UK
3Birmingham Health Partners Centre for Regulatory Science and Innovation, University of Birmingham, Birmingham, UK
4Institute of Applied Health Research, University of Birmingham, Birmingham, UK
5NIHR Surgical Reconstruction and Microbiology Research Centre, University Hospital Birmingham and University of Birmingham, Birmingham, UK
6NIHR Birmingham Biomedical Research Centre, University Hospital Birmingham and University of Birmingham, Birmingham, UK
7Centre for Trauma Science Research, University of Birmingham, UK
8Aparito Limited
9LongCOVIDSOS, UK
10Patient and Public Involvement, Therapies for Long COVID (TLC) study, University of Birmingham, UK
11Citizen Partner, COVID-END (COVID-19 Evidence synthesis Network to support Decision-making (WHO partner)
12Long COVID Scotland, UK
13DEMAND Hub, University of Birmingham, Birmingham, UK
14Health Data Research UK, London, UK
15UK SPINE, University of Birmingham, Birmingham, UK

*Corresponding author:
E-mail: s.e.hughes@bham.ac.uk

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
ABSTRACT

Objective
To describe the development and initial validation of a novel patient-reported outcome measure of Long COVID symptom burden, the Symptom-Burden Questionnaire for Long COVID (SBQ™-LC).

Method and Findings
This multi-phase, prospective mixed-methods study took place between April and August 2021 in the United Kingdom (UK). A conceptual framework and initial item pool were developed from published systematic reviews. Further concept elicitation and content validation was undertaken with adults with lived experience (n = 13) and clinicians (n = 10), and face validity was confirmed by the Therapies for Long COVID Study Patient and Public Involvement group (n = 25). The draft SBQ™-LC was field tested by adults with self-reported Long COVID recruited via social media and international Long COVID support groups (n = 274). Thematic analysis of interview and survey transcripts established content validity and informed construction of the draft questionnaire. Rasch analysis of field test data guided item and scale refinement and provided evidence of the final SBQ™-LC’s measurement properties. The Rasch-derived SBQ™-LC is composed of 17 independent scales with promising psychometric properties. Respondents rate symptom burden during the past 7-days using a dichotomous response or 4-point rating scale. Each scale provides coverage of a different symptom domain and returns a summed raw score that may be converted to a linear (0 – 100) score. Higher scores represent higher symptom burden.

Conclusions
The SBQ™-LC is a comprehensive patient-reported assessment of Long COVID symptom burden developed using modern psychometric methods. It measures symptoms of Long COVID important to individuals with lived experience and may be used to evaluate the impact of interventions and inform best practice in clinical management.
INTRODUCTION

Since its emergence in 2019, the global COVID-19 pandemic has caused more than 250 million infections and over five million deaths [1]. For many individuals, the infection is mild and short-lived; however, a proportion of those infected continue to experience, or go on to develop, long-term symptoms that persist beyond the acute phase of infection. These persistent symptoms are known collectively as “post-acute sequelae of COVID-19 (PACS)”, “post-acute COVID-19”, “post-COVID-19 syndrome”, or “Long COVID” [2].

Symptom burden may be defined as the “subjective, quantifiable prevalence, frequency, and severity of symptoms placing a physiologic burden on patients and producing multiple negative, physical, and emotional patient responses” [3]. The symptoms reported by individuals with Long COVID are heterogenous, affecting multiple organ systems with fatigue, dyspnea, and impaired concentration being among the most prevalent symptoms [4–6]. Symptoms may be persistent, cyclical, or episodic in their presentation and can pose a significant burden for affected individuals with negative consequences for work capability, functioning, and quality of life [7–9]. There is a growing body of research on the prevalence, incidence, co-occurrence, and persistence of Long COVID signs and symptoms [4,5,7,10–13]. To date, these data have largely been collected using bespoke, cross-sectional survey tools due, in part, to the limited availability of condition-specific, validated self-report instruments.

Patient-reported outcomes (PROs) are measures of health reported directly by the patient without amendment or interpretation by a clinician or anyone else [14]. Validated PRO instruments developed specifically for Long COVID that address the complex, multi-factorial nature of the condition are needed urgently to further understanding of Long COVID symptoms and underlying pathophysiology, support best practice in the clinical management of patients, and to evaluate the safety, effectiveness, acceptability, and tolerability of interventions [15–17]. Validated PRO instruments have recently been developed to measure the global impact of Long COVID and there are several unvalidated screening tools, surveys, and questionnaires also available [18,19]. However, individuals living with Long COVID have suggested that existing self-report measures fail to capture the breadth of experienced symptoms [9,20,21]. To address the need for a comprehensive measure of self-reported symptom
burden specific to Long COVID, this study used Rasch analysis to develop and validate, in accordance with US Food and Drug Administration (FDA) guidance, a novel PRO instrument, the Symptom Burden Questionnaire™ for Long COVID (SBQ™-LC) [14,22].

METHODS

Setting and study design

This multi-phase, prospective mixed-methods study (Fig 1) was nested within the Therapies for Long COVID for non-hospitalized individuals (TLC) study (ISRCTN Reference: 15674970) and took place from 14 April to 1 August 2021 [23]. A favorable ethical opinion was granted by the University of Birmingham Research Ethics Committee (Reference: ERN_21-0191).

Figure 1. Study flow diagram showing the development of SBQ™-LC (Version 1.0)
Study population

Content validation was undertaken with adults with Long COVID recruited from the TLC study’s Patient and Public Involvement and Engagement (PPIE) group and clinicians recruited from the TLC study team and Long COVID research studies based in the United Kingdom. The field test population included adults with self-reported Long COVID. Participants were aged 18 years or older who could self-complete the SBQ™-LC in English. No exclusion criteria relating to duration of Long COVID symptoms, hospitalization for SARS CoV-2, or vaccination status were applied. A minimum sample size of 250 respondents was prespecified for field testing. In Rasch analysis, a sample of 250 respondents provides 99% confidence that item calibrations and person measures are stable within ± 0.50 logits [24].

Symptom coverage and existing PROs

The SBQ™-LC’s conceptual framework was developed from systematic literature reviews of Long COVID symptoms [4,25]. Existing symptom measures (n = 6) with good face validity in the context of Long COVID were reviewed to establish whether a new PRO instrument of symptom burden was needed [26–31]. When mapped to the conceptual framework, symptom coverage of these measures ranged from 25.0% to 64.1% (mean = 42.4% SD 16.8). This finding suggested complete coverage of Long COVID symptoms could not be guaranteed using existing measures, providing justification for the development of the SBQ™-LC.

Study Procedures

Content validation

Content validation involved cognitive debriefing with adults with Long COVID and an online clinician survey. Cognitive debriefing interviews to ascertain the relevance, comprehensiveness, comprehensibility, and acceptability of the SBQ™-LC’s items for adults with lived experience of Long COVID were conducted via videoconferencing and recorded. The clinician survey was administered using the survey software application, SmartSurvey [32]. Verbatim transcripts of the interview recordings, field notes, and free text comments from the clinician survey were analyzed qualitatively using thematic analysis and a pre-specified framework to identify problems with the relevance, comprehensiveness, clarity, and acceptability of the items [33,34]. Participants with lived experience
of Long COVID were asked to identify additional symptoms not captured by the initial item pool. Findings informed revisions which were tracked per item using an Excel spreadsheet. A draft version of the SBQ™-LC was constructed and sent forward for field testing.

Field testing

The Aparito Atom5™ platform is a regulated (ISO13485, ISO/IEC 27001:2013 Accreditation, FDA CFR21 Part 11 compliant) software platform that provides remote data collection and real-time patient monitoring via a smartphone application and integration with wearable devices [35]. The draft SBQ™-LC and a demographic questionnaire were programmed onto Atom5™ for delivery as an electronic patient-reported outcome (ePRO). Participants were recruited via social media advertisements posted on Twitter and Facebook by the study team and through other support group platforms and website registrations via collaborating Long COVID support groups based in the UK. Interested individuals clicked on a URL connecting them to a study-specific website where they could read detailed information about the study, provide their informed consent, and download the Atom5™ app to their mobile device (i.e., smartphone or tablet). Participants accessed the SBQ™-LC and demographic questionnaire via a unique QR code and once the completed questionnaires were submitted, participants could delete the app from their phone. The anonymised field test data were downloaded securely from Atom5™ for analysis.

Statistical Analyses

STATA (Version 16) was used to clean and prepare the data and SPSS (Version 28.0) was used for descriptive data analyses. Rasch analysis was conducted on the field test data to refine the SBQ™-LC and assess its psychometric properties. A Rasch analysis is the formal evaluation of a PRO instrument against the Rasch measurement model. The Rasch model is a mathematical ideal that specifies a set of criteria for the construction of interval-level measures from ordinal data [36]. It is a probabilistic model that specifies that a person’s response to an item is only governed by the person and the location of the item on a shared scale measuring the latent trait. The probability that a person will endorse an item is a logistic function of the difference between a person’s trait level (expressed as person ability) and the amount of trait expressed by the item (expressed as item difficulty) [37]. Rasch analysis enabled the SBQ™-LC to be constructed as a modular PRO instrument (i.e., a multi-domain item bank) with linear, interval-level measurement properties. These properties render Rasch-
developed PROs suitable for use with individual patients as well as for group-level comparisons, permit direct comparisons of scores across domains, and facilitate the construction of alternative test formats (i.e., short forms and computer adaptive tests (CATs) [38,39].

Rasch analyses were carried out using Winsteps software (Version 5.0.5) and the Partial Credit Model (PCM) for polytomous data. The PCM was selected as the question wording and rating scale categories varied across items. Joint Maximum Likelihood Estimation (JMLE) in Winsteps enabled parameter estimation where data were missing [40]. Items scored as “not applicable” were considered ‘missing’ and did not contribute to item estimates. Misfitting response patterns (e.g., arising from respondent guessing or other unexpected behaviour) have been shown to result in biased item estimates with detrimental impacts for model fit [41]. Therefore, as is customary in Rasch analysis, person fit statistics were appraised and persons with misfitting response patterns (i.e., outfit MnSq logits > 2.0 logits) were removed iteratively and item parameters re-estimated until evidence of item parameter stability was observed [37,42].

Rating scale functioning for individual items was assessed against the following criteria: 1) items oriented in the same direction as a check for data entry errors (i.e., appraisal of point-measure correlations), 2) average category measures advance (i.e., higher categories reflect higher measures; 3) category outfit mean-square (MnSq) values do not exceed 2.0 logits (i.e., as an indicator of unexpected randomness in the model); and 4) each category endorsed by a minimum of ten respondents [43]. If an item’s rating scale failed to meet these criteria, adjacent categories were combined or the item was removed. Category probability curves (CPCs) provided a graphical representation as further evidence of rating scale functioning.

To confirm model fit, Rasch analyses (including appraisal of unidimensionality, local independence, and individual item fit statistics) were completed iteratively as items were removed or grouped to create new scales. Person reliability and separation indices and scale-to-sample targeting were also evaluated. Targeting examines the correspondence between items and persons and, for a well-targeted scale, the items in a scale should be spaced evenly across a reasonable range of the scale and correspond to the range of the construct experienced by the sample. Person reliability examines
the reproducibility of relative measure location and person separation provides a measure of the number of distinct levels of person ability (i.e., symptom burden) that can be distinguished by a scale [44]. Cronbach’s alpha, as a measure of the internal consistency reliability, was computed for each scale [45]. The list of parameters evaluated in the development and validation of the SBQ™-LC, along with acceptability criteria, are described in Table 1.
Table 1. Rasch measurement properties and acceptability criteria

<table>
<thead>
<tr>
<th>Measurement property</th>
<th>Definition/Aim of evaluation</th>
<th>Acceptability criteria for fit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Valid measurement model</strong></td>
<td>To identify a set of items which effectively measure the target construct of symptom burden in Long COVID (i.e., fulfill the axioms of fundamental measurement permitting the construction of interval level scales).</td>
<td>• Unidimensionality: PCAR(^a) highest eigenvalue of first residual contrast &lt; 2.0; disattenuated correlations &gt; 0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Local item independence: Residual correlation &lt; 0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Fit statistics: Outfit/infit MnSq values within 0.5 – 1.5 logits</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Point measure correlations &gt; 0.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Rating scales: scale oriented with latent variable, categories advance monotonically, category fit statistics: Outfit MnSq values &lt; 2.0 logits, uniform category endorsement.</td>
</tr>
<tr>
<td><strong>Reliability</strong></td>
<td>Rasch-based reliability is the share of true variance of the total observed variance of the measure.</td>
<td>• Rasch-based reliability: ( r \geq 0.70 ) acceptable ( r \geq 0.80 ) good ( r \geq 0.90 ) excellent</td>
</tr>
<tr>
<td>Rasch reliability</td>
<td>Person separation index (PSI) is the number of distinct levels of person ability (i.e., symptom burden) that can be identified by the measure.</td>
<td>( \text{PSI}^b = 1.5 – 2.0 ) acceptable level of separation ( \text{PSI} = 2.0 – 3.0 ) good level of separation</td>
</tr>
<tr>
<td></td>
<td>The extent to which items comprising a scale measure the same construct (i.e., the degree of homogeneity or relatedness among the items of a scale).</td>
<td>( \text{PSI} \geq 3.0 ) excellent level of separation</td>
</tr>
<tr>
<td>Internal consistency reliability</td>
<td></td>
<td>• Cronbach’s alpha range ( \geq 0.70 )</td>
</tr>
<tr>
<td><strong>Targeting</strong></td>
<td>The extent to which the range of the variable measured by the scale matches the range of that variable in the study sample.</td>
<td>• Item-person map</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Acceptable targeting is demonstrated by close correspondence of the person mean with the item mean for a scale (i.e., ( \pm 1.0 ) logits from the mean of zero).</td>
</tr>
</tbody>
</table>

\(^a\)Principal component analysis of the residuals; \(^b\)PSI = person separation index
RESULTS

Item development and content validation

An initial item pool (n = 97) was constructed, guided by the conceptual framework developed from the published literature. Clinician review of the item pool informed changes to improve item clarity. Content validity was confirmed by individuals with lived experience of Long COVID (n = 13) in two rounds of cognitive debriefing interviews. Participants were of white ethnicity, ten (76.9%) were female, and ranged in age from 20 – 60 years. Cognitive debriefing identified gaps in symptom coverage, resulting in the generation of 69 new items. Thematic analysis classified problems with draft items’ relevance, comprehensiveness, clarity, acceptability. S1 Table presents key themes, together with exemplar quotations, from the thematic analysis.

The draft SBQ™-LC included 166 items (155 symptoms + 11 interference items) and an a priori theoretical domain classification comprised of 14 domains, each constructed as an independent scale. Items utilized a 7-day recall period and burden was measured using a dichotomous “yes/no” response or a 5-point rating scale measuring severity, frequency, or interference. Higher scores represented greater symptom burden. Face validity of the draft SBQ™-LC was confirmed by TLC study’s PPI group.

Readability

Readability, measured using Flesch-Kincaid Reading Grade Level Test and the SMOG Index score, was calculated using the web-based application Readable [46–48]. The American Medical Association (AMA) and the National Institutes of Health (NIH) recommend readability of patient materials should not exceed a sixth-grade reading level (a Flesh-Kincaid Reading Grade Level of score of 6.0) [49]. The SBQ™-LC’s Flesch-Kincaid Grade Level score was 5.33 and SMOG index score was 8.27. Text with a SMOG index score ranging between 7 and 9 would be understood by 93% of adults in the UK [50].

Field testing

Participants
Over a two-week period, 906 questionnaires were delivered and 330 responses received. Fifty-six submissions were incomplete and excluded from the analyses. The final sample included 274 complete responses, a response rate of 30.2%. Respondents’ ages ranged from 21 to 70 years, with a mean (SD) age of 45.1 (10.1) years. The sample included 240 (87.6%) females and 253 (92.3%) respondents were white. Two hundred and eleven (77.0%) participants reported three or more comorbidities. One hundred and twenty-nine (47.1%) participants reported a diagnosis of SARS-CoV-2 that was confirmed by a positive PCR test and 22 (8.0%) respondents reported confirmation of SARS-CoV-2 by a positive lateral flow test. Eleven participants (4.0%) had been hospitalized due to SARS-CoV-2 and 70 (25.5%) had received one dose of vaccine while 187 (68.2%) had received two doses. In total, 153 (55.8%) respondents were in either full-time or part-time employment (Table 2).
Table 2. Demographic characteristics of the field test sample (n = 274)

<table>
<thead>
<tr>
<th>Category</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (n = 263 reported)</strong></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>45.1 years</td>
</tr>
<tr>
<td>SD</td>
<td>10.1 years</td>
</tr>
<tr>
<td>Range</td>
<td>21 – 70 years</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>240 (87.6)</td>
</tr>
<tr>
<td>Male</td>
<td>34 (12.4)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
</tr>
<tr>
<td>Asian or Asian British</td>
<td>7 (2.6)</td>
</tr>
<tr>
<td>Black, African, Caribbean, or Black British</td>
<td>3 (1.1)</td>
</tr>
<tr>
<td>Mixed or Multiple ethnic groups</td>
<td>11 (4.0)</td>
</tr>
<tr>
<td>Other ethnic group</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>White</td>
<td>253 (92.3)</td>
</tr>
<tr>
<td><strong>Occupational status</strong></td>
<td></td>
</tr>
<tr>
<td>Caregiver</td>
<td>3 (1.1)</td>
</tr>
<tr>
<td>Employed full-time</td>
<td>114 (41.6)</td>
</tr>
<tr>
<td>Employed but currently not working</td>
<td>51 (18.6)</td>
</tr>
<tr>
<td>Employed part-time</td>
<td>39 (14.2)</td>
</tr>
<tr>
<td>Furloughed</td>
<td>7 (2.6)</td>
</tr>
<tr>
<td>In full-time education</td>
<td>3 (1.1)</td>
</tr>
<tr>
<td>Retired</td>
<td>6 (2.2)</td>
</tr>
<tr>
<td>Voluntary work</td>
<td>3 (1.1)</td>
</tr>
<tr>
<td>Unemployed</td>
<td>20 (7.3)</td>
</tr>
<tr>
<td>Other</td>
<td>28 (10.2)</td>
</tr>
<tr>
<td><strong>Positive PCR Test for SARS-CoV-2 infection</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>129 (47.1)</td>
</tr>
<tr>
<td>No</td>
<td>137 (50.0)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>8 (2.9)</td>
</tr>
<tr>
<td><strong>Positive lateral flow test for SARS-CoV-2 infection</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>22 (8.0)</td>
</tr>
<tr>
<td>No</td>
<td>227 (82.8)</td>
</tr>
<tr>
<td>Don’t know</td>
<td>25 (9.1)</td>
</tr>
<tr>
<td><strong>Hospitalized for SARS-CoV-2 infection</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>11 (4.0)</td>
</tr>
<tr>
<td>No</td>
<td>263 (96.0)</td>
</tr>
<tr>
<td><strong>Admitted to intensive care unit due to SARS-CoV-2 infection</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>No</td>
<td>274 (100.0)</td>
</tr>
<tr>
<td><strong>Attended hospital emergency department due to SARS-CoV-2 infection</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>111 (40.5)</td>
</tr>
<tr>
<td>No</td>
<td>163 (59.5)</td>
</tr>
<tr>
<td><strong>Vaccine status as of (XX-Jun-2021)</strong></td>
<td></td>
</tr>
<tr>
<td>Two doses</td>
<td>187 (68.2)</td>
</tr>
<tr>
<td>One dose</td>
<td>70 (25.5)</td>
</tr>
<tr>
<td>No dose</td>
<td>17 (6.2)</td>
</tr>
<tr>
<td><strong>Received shielding letter from UK government</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>12 (4.4)</td>
</tr>
<tr>
<td>No</td>
<td>262 (95.6)</td>
</tr>
<tr>
<td><strong>Care home resident</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>3 (1.1)</td>
</tr>
<tr>
<td>No</td>
<td>271 (98.9)</td>
</tr>
<tr>
<td><strong>EQ5D utility index score (n = 262)</strong></td>
<td>0.490 (SD 0.260)</td>
</tr>
</tbody>
</table>

*aShielding letter = letter to indicate clinical vulnerability requiring enhanced social distancing.*
Assessment of rating scale function, item and scale refinement

Rating scale functioning was assessed by examining the relevant Winsteps output tables and item category probability curves (S1 Figure). Appraisal of category endorsement revealed the presence of floor effects and a positively skewed distribution. To address non-uniform category distribution, the 5-point rating scale was collapsed to either a dichotomous response or 4-point rating scale (where 0 = none, 1 = mild, 2 = moderate, 3 = severe). After rating scale adjustment, point-measure correlations were found to be positive (range = 0.23 - 0.92), categories were ordered, and category fit statistics were productive for measurement (Outfit MnSq < 2.0 logits) for all items. Category distribution remained positively skewed (mean skewness = 1.44, mean SE = 0.15, range = -2.30 - 11.64). Disordered thresholds were observed for 52 (39.7%) items. Threshold disordering is indicative of a category occupying a narrow interval on the rating scale and is only considered a cause for concern if category disordering is also observed [40]. In total, 35 items were removed (Fig 1 shows the process of item reduction and reasons for removal). The remaining 131 items were systematically grouped to construct scales that were clinically sensible and satisfied the Rasch model.

Calibration of the final SBQ™-LC

Following response scale optimization and scale refinement, Rasch analyses were undertaken to report the psychometric properties of the final SBQ™-LC Version 1.0 (Table 3). The SBQ™-LC Version 1.0 is composed of 17 independent scales (Fig 2). All scales met the Rasch model criteria for unidimensionality and item fit. The first residual contrast values from principal component analyses of the residuals (PCAR) ranged from 1.46 to 2.03 eigenvalue units. No serious misfit was identified. Item infit MnSq values ranged from 0.67 to 1.32 logits and outfit MnSq values ranged from 0.44 to 1.53 logits. Fourteen item pairs showed local item dependency with residual correlation values > 0.4 (range = 0.44 – 0.88). In practical terms, a degree of local dependency is always observed in empirical data; therefore, it is necessary to consider the implications of item removal [51]. Despite evidence of local item dependence for eight scales, these items were retained to ensure comprehensive symptom coverage.
<table>
<thead>
<tr>
<th>Scale (Symptom domain)</th>
<th>No items</th>
<th>Misfitting items</th>
<th>Persons with misfitting responses removed (%)</th>
<th>Mean person ability (logits)</th>
<th>PCA residual eigenvalue (1st contrast)$^a$</th>
<th>Dependent item pairs</th>
<th>Item separation</th>
<th>Item reliability</th>
<th>Person separation$^b$</th>
<th>Person reliability$^c$</th>
<th>Internal consistency reliability$^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breathing</td>
<td>7</td>
<td>0</td>
<td>31 (11.3)</td>
<td>-1.95 (SE 0.99)</td>
<td>1.72</td>
<td>0</td>
<td>7.68</td>
<td>0.98</td>
<td>1.8</td>
<td>0.76</td>
<td>0.84</td>
</tr>
<tr>
<td>Circulation</td>
<td>5</td>
<td>0</td>
<td>17 (6.2)</td>
<td>-1.16 (SE 0.98)</td>
<td>1.58</td>
<td>0</td>
<td>7.43</td>
<td>0.98</td>
<td>1.1</td>
<td>0.55</td>
<td>0.63</td>
</tr>
<tr>
<td>Fatigue</td>
<td>4</td>
<td>0</td>
<td>33 (12.0)</td>
<td>3.98 (SE 1.45)</td>
<td>1.46</td>
<td>1</td>
<td>1.34</td>
<td>0.64</td>
<td>1.78</td>
<td>0.76</td>
<td>0.91</td>
</tr>
<tr>
<td>Memory, Thinking &amp; Communication</td>
<td>10</td>
<td>0</td>
<td>12 (4.38)</td>
<td>-1.04 (SE 0.54)</td>
<td>1.76</td>
<td>0</td>
<td>12.22</td>
<td>0.99</td>
<td>2.58</td>
<td>0.87</td>
<td>0.9</td>
</tr>
<tr>
<td>Sleep</td>
<td>4</td>
<td>0</td>
<td>31 (11.3)</td>
<td>-0.10 (SE 0.78)</td>
<td>1.62</td>
<td>2</td>
<td>13.44</td>
<td>0.99</td>
<td>1.67</td>
<td>0.74</td>
<td>0.56</td>
</tr>
<tr>
<td>Movement</td>
<td>3</td>
<td>0</td>
<td>31 (11.3)</td>
<td>-6.27 (SE 2.06)</td>
<td>1.57</td>
<td>3</td>
<td>11.23</td>
<td>0.99</td>
<td>2.08</td>
<td>0.81</td>
<td>0.86</td>
</tr>
<tr>
<td>Muscles &amp; Joints</td>
<td>9</td>
<td>0</td>
<td>23 (8.39)</td>
<td>-0.85 (SE 0.54)</td>
<td>1.68</td>
<td>0</td>
<td>7.38</td>
<td>0.98</td>
<td>1.68</td>
<td>0.73</td>
<td>0.84</td>
</tr>
<tr>
<td>Skin &amp; Hair</td>
<td>8</td>
<td>0</td>
<td>26 (9.49)</td>
<td>-2.38 (SE 0.92)</td>
<td>1.55</td>
<td>0</td>
<td>4.6</td>
<td>0.95</td>
<td>0.71</td>
<td>0.34</td>
<td>0.68</td>
</tr>
<tr>
<td>Eyes</td>
<td>10</td>
<td>0</td>
<td>12 (4.38)</td>
<td>-1.21 (SE 0.72)</td>
<td>1.85</td>
<td>0</td>
<td>4.28</td>
<td>0.95</td>
<td>1.08</td>
<td>0.54</td>
<td>0.72</td>
</tr>
<tr>
<td>Ear, Nose &amp; Throat</td>
<td>14</td>
<td>0</td>
<td>11 (4.01)</td>
<td>-1.25 (SE 0.49)</td>
<td>1.96</td>
<td>1</td>
<td>3.39</td>
<td>0.92</td>
<td>1.22</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Stomach &amp; Digestion</td>
<td>8</td>
<td>1</td>
<td>23 (8.39)</td>
<td>-1.52 (SE 0.70)</td>
<td>1.6</td>
<td>0</td>
<td>6.19</td>
<td>0.97</td>
<td>0.95</td>
<td>0.48</td>
<td>0.75</td>
</tr>
<tr>
<td>Mental Health &amp; Wellbeing</td>
<td>9</td>
<td>0</td>
<td>25 (9.12)</td>
<td>-0.70 (SE 0.52)</td>
<td>1.66</td>
<td>0</td>
<td>8.71</td>
<td>0.99</td>
<td>1.78</td>
<td>0.76</td>
<td>0.82</td>
</tr>
<tr>
<td>Female Reproductive &amp; Sexual Health</td>
<td>7</td>
<td>0</td>
<td>27 (9.85)</td>
<td>-2.07 (SE 1.05)</td>
<td>1.99</td>
<td>2</td>
<td>5.52</td>
<td>0.97</td>
<td>0.99</td>
<td>0.5</td>
<td>0.59</td>
</tr>
<tr>
<td>Male Reproductive &amp; Sexual Health</td>
<td>3</td>
<td>0</td>
<td>1 (3.03)</td>
<td>-1.48 (SE 1.98)</td>
<td>2.03</td>
<td>2</td>
<td>2.27</td>
<td>0.84</td>
<td>0.79</td>
<td>0.38</td>
<td>0.6</td>
</tr>
<tr>
<td>Pain</td>
<td>4</td>
<td>0</td>
<td>25 (9.12)</td>
<td>-1.53 (SE 1.05)</td>
<td>1.75</td>
<td>2</td>
<td>13.06</td>
<td>0.99</td>
<td>1.62</td>
<td>0.72</td>
<td>0.77</td>
</tr>
<tr>
<td>Other symptoms</td>
<td>18</td>
<td>0</td>
<td>22 (8.03)</td>
<td>-1.53 (SE 0.43)</td>
<td>1.67</td>
<td>1</td>
<td>5.5</td>
<td>0.97</td>
<td>1.33</td>
<td>0.64</td>
<td>0.79</td>
</tr>
<tr>
<td>Interference</td>
<td>8</td>
<td>0</td>
<td>21 (7.66)</td>
<td>2.35 (SE 0.89)</td>
<td>1.95</td>
<td>0</td>
<td>11.21</td>
<td>0.99</td>
<td>2.13</td>
<td>0.82</td>
<td>0.89</td>
</tr>
</tbody>
</table>

$^a$ Unidimensionality: PCAR eigenvalue of first residual contrast < 2.0; Person separation: ≥1.50 acceptable, ≥2.00 good, ≥3.00 excellent; $^b$ Rasch-based reliability: $r \geq 0.70$ acceptable, $r \geq 0.80$ good, $r \geq 0.90$ excellent; $^c$ Cronbach’s $\alpha \geq 0.70$ acceptable
Lastly, scale-to sample targeting and person reliability and separation were evaluated. Four scales had mean person ability values within ±1.0 logits of mean item difficulty. Mean person ability ranged from -6.27 logits to 3.98 logits (SE range = 0.43 - 2.06). S2 Figure shows the item-person maps for the SBQ™-LC’s scales. Person reliability ranged from 0.34 to 0.87 and person separation indices ranged from 0.71 – 2.58. Values for Cronbach’s alpha, as evidence of internal consistency reliability, ranged from 0.56 to 0.91.
DISCUSSION

This study described the development and initial validation of the SBQ™-LC, a Rasch-developed, modular PRO instrument constructed to comprehensively assess symptom burden in Long COVID. The SBQ™-LC was developed in accordance with international, consensus-based standards and regulatory guidance and may be used to evaluate the impact of interventions and to inform clinical care [14,22,52]. The findings from published systematic reviews were used to construct a conceptual framework and generate an initial item pool. Rigorous content validity testing provided evidence of the SBQ™-LC’s relevance, comprehensiveness, comprehensibility, and acceptability. Rasch analysis permitted optimisation of the SBQ™-LC’s items and response scales.

The SBQ™-LC was developed with the extensive involvement of adults with lived experience of Long COVID and is a strength of this study. Involvement of the target population is regarded as the “gold standard” in the development of PRO instruments. Patient involvement may be considered of particular importance in the context of Long COVID where there is a rapidly evolving evidence-base and affected individuals have reported experiences of stigma and a lack of acknowledgement from the medical community regarding the breadth and nature of their symptoms [9]. Involvement of adults with Long COVID in all study phases (i.e., development, refinement, and validation) ensured the patient perspective was embodied in the SBQ™-LC’s items.

Rasch analysis of the SBQ™-LC guided item reduction and rating scale refinement, optimised measurement accuracy, and minimized respondent burden. These analyses showed several of the SBQ™-LC’s scales to be off-target with evidence of low person reliability and separation. Poor scale-to-sample targeting is indicative of the items within a scale failing to provide full coverage of person locations (i.e., range of symptom burden experienced by the sample) [53]. Negative mean person measures (>1.0 logits), floor effects, and positively skewed response category distributions suggested the SBQ™-LC may be targeting individuals with higher levels of symptom burden than the level of burden represented by the field test sample. Highly skewed samples and poor targeting can produce low person reliability coefficients even if an instrument is functioning as intended and provides a possible explanation for the low person reliability values observed in this study [53].
As a Rasch-developed instrument, the SBQ™-LC’s ordinal raw scales may be converted to linear scales, with each 1-point change in a scale score being equidistant across the entire scale. Linear scores will enable the direct comparison of scores across the symptom domains represented by the SBQ™-LC’s scales. This modular construction means researchers and clinicians need only select those scales required to provide targeted assessment of a particular symptom domain, reducing respondent burden further by removing the need to complete the SBQ™-LC in its entirety. Moreover, the Rasch model will make it possible to compare data from the SBQ™-LC with other PROs via co-calibration studies. As each item of a Rasch-derived scale functions independently from others on that scale, the SBQ™-LC can be adapted for administration as a computer adaptive test (CAT) [39,54]. A CAT is a test, administered via a computer, which adapts to the respondent’s ability in real-time by selecting different questions from an item bank to provide a more accurate measure of the respondent’s ability without the need to administer a large number of items [55]. CATs can reduce respondent burden, improve accuracy, and provide individualized assessment, instrument characteristics which are attractive when assessing a health condition with heterogeneous, relapsing and remitting symptoms such as Long COVID.

The burden of Long COVID on healthcare systems continues to grow as more individuals are infected by SARS-CoV-2 [56]. To meet this growing demand, services require cost-effective resources to support safe, effective clinical management. The use of the SBQ™-LC in the TLC study will provide early evidence of feasibility of the SBQ™-LC for use in remote patient monitoring [57]. A previous randomized controlled trial has shown remote symptom monitoring using PROs can result in fewer emergency department attendances, reduce hospital admissions, prompt earlier intervention, and improve patients’ health-related quality of life [58]. If used in a clinical trial, symptom data collected remotely via the SBQ™-LC could provide valuable information on the safety, efficacy, and tolerability of new interventions for Long COVID [15]. If used within routine care, the SBQ™-LC has potential to facilitate patient-clinician conversations, guide treatment decision-making, and facilitate referrals to specialist services [58–60].
The use of social media to recruit participants meant it was not possible to confirm representativeness of the field test sample and is a limitation of this study. The relatively low completion rate suggested potential field test participants (i.e., possibly individuals experiencing higher levels of symptom burden, for example, symptoms such as fatigue or cognitive dysfunction) may have been deterred by the effort required to complete the SBQ™-LC or lacked sufficient incentive to participate. As such, the level of symptom burden captured in the field test may not reflect accurately the wider experience of Long COVID patients. Lastly, despite attempts to sample purposively for maximum variation, the demographics of the content validation study sample were highly skewed towards white (100%) females (76.9%). Further validation, undertaken as part of the TLC study which will prospectively target higher representation from ethnic minority groups, is planned and aims to confirm these findings. Further research is also needed to evaluate the SBQ™-LC using traditional psychometric indicators including test-retest reliability, construct validity, responsiveness, and measurement error. Studies to explore the feasibility and acceptability of the SBQ™-LC for use in clinical settings are also needed. The SBQ™-LC is currently available in UK English as an ePRO and in paper form. Linguistic and cross-cultural validation studies will ensure its generalizability outside of the UK and make translated versions of the SBQ™-LC available for use.

The presence of symptoms of COVID-19 persisting beyond the acute phase of infection in a significant number of patients represents an ongoing challenge for healthcare systems globally. High quality PRO instruments with robust psychometric properties are required to better understand the signs, symptoms, and underlying pathophysiology of Long COVID, to develop safe and effective interventions, and to meet the day-to-day needs of this growing patient group. The SBQ™-LC was developed as a comprehensive measure of Long COVID symptom burden. With promising psychometric properties, the SBQ™-LC may be suitable for use in Long COVID research studies and in the delivery of clinical care.

Acknowledgements:
The authors gratefully acknowledge the contributions of the clinicians who participated in the online survey. We thank Dr Anita Slade for their support with interpretation of the Rasch analyses.

The SBQ™-LC Version 1.0 is available under license. To request a copy, please e-mail: tlcstudy@contacts.bham.ac.uk.

Contributors:
Concept and design: Calvert, Haroon, Hughes, Aiyegbusi, Frost, Davies
Acquisition, analysis, or interpretation of data: Hughes, Haroon, Calvert, Aiyegbusi, Matthews,
Camaradou, Price, Ormerod, McMullan, Davies, Frost, McNamara

Drafting of the manuscript: Hughes, Calvert, Haroon, Submaranian

Critical revision of the manuscript: Hughes, Calvert, Haroon, Submaranian, McMullan, Aiyegbusi,
Davies, Frost, McMullan, Davies, Matthews, Camaradou, Jackson, Turner, Price, Ormerod, Walker

Statistical analysis: Hughes, Submaranian, Jackson

Obtained funding: Haroon, Calvert, Hughes, Aiyegbusi, McMullan, Davies

Administration, technical, or material support: Frost, Davies, Submaranian, Jackson, McNamara

Walker

Competing interests: SEH receives funding from the National Institute for Health Research (NIHR)
Applied Research Collaboration (ARC) West Midlands, UK Research and Innovation (UKRI) and
declares personal fees from Aparito Limited and Cochlear Limited outside the submitted work. MJC is
Director of the Birmingham Health Partners Centre for Regulatory Science and Innovation, Director of
the Centre for the Centre for Patient Reported Outcomes Research and is a National Institute for
Health Research (NIHR) Senior Investigator. MJC receives funding from the National Institute for
Health Research (NIHR), UK Research and Innovation (UKRI), NIHR Birmingham Biomedical
Research Centre, the NIHR Surgical Reconstruction and Microbiology Research Centre, NIHR ARC
West Midlands, UK SPINE, European Regional Development Fund – Demand Hub and Health Data
Research UK at the University of Birmingham and University Hospitals Birmingham NHS Foundation
Trust, Innovate UK (part of UKRI), Macmillan Cancer Support, UCB Pharma, Janssen, GSK and
Gilead. MJC has received personal fees from Astellas, Aparito Ltd, CIS Oncology, Takeda, Merck,
Daichi Sankyo, Glaukos, GSK and the Patient-Centered Outcomes Research Institute (PCORI)
outside the submitted work. In addition, a family member owns shares in GSK. OLA receives funding
from the NIHR Birmingham ARC, West Midlands, Birmingham, NIHR ARC West Midlands, UKRI,
Health Foundation, Janssen, Gilead and GSK. He declares personal fees from Gilead Sciences Ltd,
Merck and GSK outside the submitted work. JC is a lay member on the UK NICE COVID expert
panel, a citizen partner to the COVID END Evidence Synthesis Network, PPI lead on the NIHR
CICADA ME Study, patient representative at the EAN European Neurology Autonomic Nervous
Systems Disorders Working Group, a member of the MRC/UKRI Advanced Pain Discovery Platform,
and external board member of Plymouth Institute of Health. She also reports contracts with GSK and
Medable. CM receives funding from NIHR SRMRC, UKRI, and declares personal fees from Aparito
Ltd. KM is employed by the NIHR. SH receives funding from NIHR and UKRI. No other disclosures
were reported.

Funding/Support: This work is independent research jointly funded by the National Institute for
Health Research (NIHR) and UK Research and Innovation (UKRI) (Therapies for Long COVID in non-
hospitalized individuals: From symptoms, patient reported outcomes and immunology to targeted
therapies (The TLC Study), (COV-LT-0013). The views expressed in this publication are those of the
author(s) and not necessarily those of the NIHR, the Department of Health and Social Care, or UKRI.

Role of the Funders/Sponsors: The funders had no role in the design and conduct of the study
including: the collection, management, analysis, and interpretation of the data; preparation and review
of the manuscript.

Data availability statement: Data for this project are not currently available for access outside the
TLC research team. The dataset may be shared when finalised but this will require an application
to the data controllers. The data may then be released to a specific research team for a specific
project dependent on the independent approvals being in place.

REFERENCES

1. COVID-19 situation update worldwide, as of week 44, updated 11 November 2021. [cited 16 Nov
cases

the long-term effects of COVID-19. [cited 3 Dec 2021]. Available from:
https://app.magicapp.org/#/guideline/EQpzKn/section/n3vwoL


31. Smart Survey. Available from: https://www.smartsurvey.co.uk/


47. Readable. Available: www.readable.io


SUPPORTING INFORMATION CAPTIONS

S1 Table. Themes and exemplar quotations from content validation of the SBQ™-LC item pool (cognitive debriefing and online survey)

S1 Figure. Item category probability curves for the final SBQ™-LC (Version 1.0)
Each graph, the y-axis represents the expected probability of endorsement of any given category when a person responds to the item. The x-axis represents the person ability relative to the item difficulty, with origins set to 0. The scale is measured in logits. Locations to the right of ‘0’ represent greater and greater levels of symptom burden. Moving to the left of ‘0’ locations represent lower levels of symptom burden. Categories should advance in ascending order from left to right along the x-axis and each have a distinct peak (indicating that the category is the most probable (modal) category at that point on the latent variable). The crossover points between two curves are the equal probability points of thresholds.

S2 Figure. Item-person maps for the 17 scales of the final SBQ™-LC (Version 1.0)
For each scale, the person-item map displays the location of person abilities and item difficulties respectively along the same latent dimension (y-axis).