PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Development and Validation of a Patient Reported Measure of Compassion in Healthcare: The Sinclair Compassion Questionnaire (SCQ)
AUTHORS	Sinclair, Shane; Hack, Thomas; MacInnis, Cara; Jaggi, Priya; Boss, Harrison; McClement, Susan; Sinnarajah, Aynharan; Thompson, Genevieve

VERSION 1 – REVIEW

REVIEWER	Consedine, Nathan University of Auckland, Faculty of Medical and Health Sciences
REVIEW RETURNED	15-Dec-2020

GENERAL COMMENTS	The submitted report presents data from a pair of instrument development studies, describing the development of a measure of the patient's experience of compassion. As part of a long-term program of research in a critically under-studied area, the attempt to develop a measure of the patient's experience of compassion is important and likely to be of broad interest in the field. Although there are some areas in which interpretations might be more circumspect or (conversely) greater detail would be useful, the research has been carefully conducted and has the potential to make a useful contribution. There are, however, a few interrelated theoretical, methodological, and operational issues that detract from the work as it currently stands. These issues are described in greater detail below. First, my sense is that the underlying construct needs to be described differently/more precisely and/or feel that "compassion" is not particularly illuminating as description of the base content area; it looks like the intention is to assess "the patient experience of compassion" or the "perception of care" or something similar. Given how little we know about the extent to which such items reflect patient perceptions, patient expectations or models of what care "looks like" (much less whether they match those of providers), I think calling this a measure of compassion is imprecise. While I am comfortable with the labelling of the scale, I think it should be discussed in the text as a more precise construct than a measure of "compassion". As these authors are aware through my review of the prior "grounded theory" paper (see https://bmjopen.bmj.com/content/8/3/e019701), I have some reservations about treating data reflecting the views, attitudes, opinions, and feelings of what I see as rather particular samples (from palliative, LT care, etc) as necessarily indicative of broader processes of patient or professional samples in healthcare; the

	submission briefly acknowledges this issue on p. 20. While I see no problem in the basic face validity of the items used in the construction of this scale or in the methods (although see below), I do wonder whether the expectations, attitudes, feelings, needs, etc of persons in terminal care contexts might systematically vary in important ways (e.g., prioritizing kindnesses over care or vice versa). This concern is compounded by the c. 45% response rate in the two studies. Although this is not a bad rate and the mere fact that it actually quantified is a strength of the study, it seems likely that a particular and self-selected sample who were already "special" by virtue of the recruitment context (indexing age, health, time horizon, etc) may or may not reflect more general views. Combined with the absence of a strong conceptual basis for these recruitment sites, I think the submission would be improved with a slightly more circumspect, restrictive, or precise interpretative mindset and a more explicit acknowledgment of this limitation. Three, my feeling is that the submission overstates the extent to which validity has been demonstrated and makes an unnecessarily general claim; the expression "highly valid" is used at several points. More fully, while I agree that the methods behind the measure offer a fair degree of confidence in both internal and test-retest reliability as well as face and convergent validity, I am less sure that clear evidence of discriminant and construct (fit to theory) validity has been provided. To date, the evidence consists of (a) 0.60 correlation with a patient experience measure and (b) minimal associations with a measure of depressive symptomology. There was no evidence that it was distinct from empathy or patient-centred care or other constructs, hence, in my opinion, the offering in the submission is better characterized as preliminary evidence. As importantly, it was not clear to me that the single component SCQ should be interpreted as being consistent with the Patient Compassion Model. There is minimal discussion of this issue – a lack of construct validity – in the manuscript as submitted. Instead, there is some general method discussion (p. 18) that obscures the fact that the measure did not, in fact, map onto the construct as conceptualized. For example, it is noted that " . . . the perspective (sic) of patients were incorporated across each stage of this study fortifying our foundational patient-centred research establishing construct validity" (p. 19). Such statements might be taken as implying the demonstration of forms of validity that have not yet been shown. Again, while the instrument is very promising and the submission provides a welcome addition to the very few tools in the area, a more circumspect treatment of these issues is needed. As noted, my feeling is that the methods employed in the study are generally appropriate and, as importantly, are transparently described. Although professional contingencies tend to favour "siloed" research conducted in isolation, studies in compassion are almost necessarily a collaborative endeavor. The authors are thus to be commended for the transparency with which the multiple psychometric decisions have been made and communicated. That said, I have a few requests for clarification: - Can the authors comment on why only N = 65 were included in the test-retest analyses and why a 24-hour window was employed? Again, I am assuming there is a reason but in the interests of future studies of this kind, knowing the reasons would be of interest. - I am not sure that test-retest is the optimal basis for the removal of items. I would have thought that maximizing content coverage

	and/or construct validity (fit to theory) would be more pressing concerns. Why is reliability over time being prioritized in this way? - The paper would benefit from clarifying what the “importance” ratings are about. Given the use of these ratings in the item reduction process, I would like to see the exact item. - The information in the tables suggests a consistent degree of negative skew in the data (as well as a persistent kurtosis), presumably because people were rating their providers as very compassionate much of the time. While this is a widespread problem in research of this kind (i.e., vis-à-vis compassion in health), there is clearly a loss of discrimination at the top/positive end of the scale. Is there a reason that social desirability was not controlled and can the authors comment on the possibility of using a more “extreme” anchor (e.g., always) to fix this problem? Other minor points: - Data were not was (multiple instances) - Please clarify in what ways this measure represents a “gold standard”.

REVIEWER	Durkin, Mark Department of Psychology, University of Bolton
REVIEW RETURNED	27-Jan-2021

GENERAL COMMENTS	This is a well thought out paper that describes the need, processes, and implications of a patient-reported scale for compassion.I really enjoyed reading this article and think that The Sinclair Compassion Questionnaire addresses the need for a valid and reliable measure of compassion that has been developed with and for patients, adding to the literature in this much-needed area. While the scale has been thoroughly tested using advanced psychometrics, I would suggest changing the wording slightly on page 4 line 10, and again on page 23 line 38, to this scale "could become the Gold Standard" compassion measure for patients, as other patient measures exist in the literature.

VERSION 1 – AUTHOR RESPONSE

Reviewer #1	Authors’ comments
The submitted report presents data from a pair of instrument development studies, describing the development of a measure of the patient’s experience of compassion. As part of a long-term program of research in a critically under-studied area, the attempt to develop a measure of the patient’s experience of compassion is important and likely to be of broad interest in the field. Although there are some areas in which interpretations might be more circumspect or (conversely) greater detail would be useful, the research has been carefully conducted and has the potential to make a useful contribution. There are,	Thank you.

however, a few interrelated theoretical, methodological, and operational issues that detract from the work as it currently stands. These issues are described in greater detail below.	Thank you for raising this issue. We have provided a point-by-point response to your feedback below.
First, my sense is that the underlying construct needs to be described differently/more precisely and/or feel that "compassion" is not particularly illuminating as description of the base content area; it looks like the intention is to assess "the patient experience of compassion" or the "perception of care" or something similar. Given how little we know about the extent to which such items reflect patient perceptions, patient expectations or models of what care "looks like" (much less whether they match those of providers), I think calling this a measure of compassion is imprecise. While I am comfortable with the labelling of the scale, I think it should be discussed in the text as a more precise construct than a measure of "compassion".	Thank you for this astute comment which we discussed in detail. We agree that this measure is more correctly a measure of patients experience of compassion and have modified this throughout. Specifically, we have been careful to refer to the measure as a patient-reported compassion measure throughout the manuscript, including instances where this was previously unclear (Background, 2nd paragraph, Sentence 1; Discussion, 3rd paragraph, Line 2 & 3; Conclusion; 2nd sentence). Additionally, we have clarified that the construct of interest is patients' experiences of compassion (Strength and Limitations of Study, Bullet 4; Discussion, 1st paragraph, 2nd Sentence; Discussion, 1st paragraph, line 1 & 3rd paragraph 1st Sentence) as we had originally intended. Finally, to provide further clarity, in discussing the limitations of existing measures we now note that we are referring to compassion measures *in healthcare* (Discussion, 2nd paragraph, 1st Sentence and the Title) specifically.

As these authors are aware through my review of the prior “grounded theory” paper (see https://bmjopen.bmj.com/content/8/3/e019701), I have some reservations about treating data reflecting the views, attitudes, opinions, and feelings of what I see as rather particular samples (from palliative, LT care, etc) as necessarily indicative of broader processes of patient or professional samples in healthcare; the submission briefly acknowledges this issue on p. 20. While I see no problem in the basic face validity of the items used in the construction of this scale or in the methods (although see below), I do wonder whether the expectations, attitudes, feelings, needs, etc of persons in terminal care contexts might systematically vary in important ways (e.g., prioritizing kindnesses over care or vice versa). This concern is compounded by the c. 45% response rate in the two studies. Although this is not a bad rate and the mere fact that it actually quantified is a strength of the study, it seems likely that a particular and self-selected sample who were already “special” by virtue of the recruitment context (indexing age, health, time horizon, etc) may or may not reflect more general views. Combined with the absence of a strong conceptual basis for these recruitment sites, I think the submission would be improved with a slightly more circumspect, restrictive, or precise interpretative mindset and a more explicit acknowledgment of this limitation.	Thank you for this query. We agree, that are patient sample was somewhat limited, and have further qualified this limitation on page 20. We also mention this at the outset of the paper (3rd bullet of the Strengths and Limitations table). In addition to making this limitation more explicit, as a research team we have also now checked for any statements in the manuscript that overstate the generalizability of the measure, modifying the manuscript accordingly. We have also clarified that our sample was not terminal/palliative care but individuals living with a life limiting illness (dementia, frail elderly, renal disease, COPD, cardiac disease, cancer) (Methods: Phase 1-EFA, 1st paragraph, 2nd sentence).

Three, my feeling is that the submission overstates the extent to which validity has been demonstrated and makes an unnecessarily general claim; the expression "highly valid" is used at several points. More fully, while I agree that the methods behind the measure offer a fair degree of confidence in both internal and test-retest reliability as well as face and convergent validity, I am less sure that clear evidence of discriminant and construct (fit to theory) validity has been provided. To date, the evidence consists of (a) 0.60 correlation with a patient experience measure and (b) minimal associations with a measure of depressive symptomology. There was no evidence that it was distinct from empathy or patient-centred care or other constructs, hence, in my opinion, the offering in the submission is better characterized as preliminary evidence. As importantly, it was not clear to me that the single component SCQ should be interpreted as being consistent with the Patient Compassion Model. There is minimal discussion of this issue – a lack of construct validity – in the manuscript as submitted. Instead, there is some general method discussion (p. 18) that obscures the fact that the measure did not, in fact, map onto the construct as conceptualized. For example, it is noted that " . . . the perspective (sic) of patients were incorporated across each stage of this study fortifying our foundational patient-centred research establishing construct validity" (p. 19). Such statements might be taken as implying the demonstration of forms of validity that have not yet been shown. Again, while the instrument is very promising and the submission provides a welcome addition to the very few tools in the area, a more circumspect treatment of these issues is needed.	Thank you. We have curbed our enthusiasm and removed incidences where we use the term 'highly' and 'gold standard', as this would require further testing and formal assessment of the SCQ against other compassion measures in healthcare. We also removed other potential overstatements about the validity of the measure (Results Phase 2, 4th paragraph, 1st sentence), including acknowledging the preliminary/initial nature of this evidence (Discussion, 1st paragraph, 1st sentence)—thank you for bringing this to our attention. In regard to the measures chosen for convergent and divergent validity testing, as a research team we were forced to make difficult decisions, that took into account the quality of the measures, feasibility and protocol burden. We selected what we felt where the best measures at the time but agree that further convergent and divergent validity testing needs to be conducted which we further qualify in the revised manuscript to reflect the preliminary nature of the evidence (Discussion, 1st paragraph, 1st sentence). We also note that evidence for discriminant validity also includes no significant relationships between the SCQ and other symptomology beyond depression and well-being measured by the ESAS-r (e.g. pain, fatigue, anxiety, sleep, etc.). We have made this explicit in the manuscript (Results: Phase 2, 4th paragraph, 7th sentence). We apologize for the lack of clarity on our part, related to the congruence between the Patient Compassion Model and the SCQ. We have clarified this further in the revised manuscript (Discussion, 1st paragraph, 2nd sentence). We believe that this modification better reflects our original intention, that the *theoretical* domains of the PCM are subsumed under a single latent construct of compassion, which is a defining feature of a reflective measure (Discussion, 1st paragraph, 3rd sentence). To expand further, Pett et al (2003) note when exploring the decision-making process for item analysis/selection, that it is important to make decisions not only based on psychometrics, but to also consider how those

	items fit together *theoretically*, specifically- "decisions should be based not only on our statistical findings but also on our knowledge about how the items fit together both rationally and theoretically" (chapter 6, p.23). For that reason, we considered the importance of our theoretical subdomain coverage for selecting final items within the context of our psychometric findings. Our decision was made in a way that would not compromise the psychometric validity of the SCQ, while taking into consideration previous theoretical work (e.g. retaining at least 2 items from each subdomain, and only selecting items based on a top-down approach by which we only selected those with highest factor loadings), and reduce the items to a more manageable length. In short, we do not believe that these theoretical subdomains detract from the single factor solution, but simply and broadly demonstrate the complexity of compassion and the foundational work on the construct conducted to date. We have added a couple of sentences to further clarify this point (Phase 1 results, 3rd paragraph, 3-4th sentences). Finally, we have also modified and tempered the referenced sentence (Discussion, 2nd paragraph, 3rd sentence).
As noted, my feeling is that the methods employed in the study are generally appropriate and, as importantly, are transparently described. Although professional contingencies tend to favour "siloed" research conducted in isolation, studies in compassion are almost necessarily a collaborative endeavor. The authors are thus to be commended for the transparency with which the multiple psychometric decisions have been made and communicated	Thank you.
That said, I have a few requests for clarification:	Please see our responses to each of these below
Can the authors comment on why only N = 65 were included in the test-retest analyses and why a 24-hour window was employed? Again, I am assuming there is a reason but in the	The n=65 is derived a priori, from formulas to assess ICC sample size from Streiner et al. 2015 (ref #22, p. 189). We also note that although the larger the number of participants for test-retest the better, exceeding "50 subjects in many situations

interests of future studies of this kind, knowing the reasons would be of interest.	is probably statistical overkill" (Steiner et al, 2015, p.192). As a research team we surmised that a 24hr period was appropriate for this patient population, taking into account that they often have rapid changes in their health status and changes in the composition of their healthcare team. While some commonly cited rules of thumb suggest test-retest periods of 1 to 2 weeks (e.g. Nunnally & Bernstein, 1994), this was deemed not-appropriate given the nature of our construct of interest and patient population (decline in cognition, symptom issues). For example, in assessing compassion delivered by the healthcare team, we also chose this test-retest period to control against changes in the composition of the patients' care team (e.g. shift RNs, shift Licensed Practical Nurses) as this would have the possibility of negatively impacting test-retest scores.
I am not sure that test-retest is the optimal basis for the removal of items. I would have thought that maximizing content coverage and/or construct validity (fit to theory) would be more pressing concerns. Why is reliability over time being prioritized in this way?	Thank you for providing the opportunity to clarify our decisions surrounding item reduction and the use of test-retest as an additional means for item reduction. Given the unidimensional nature of our scale, we did not have opportunity to use commonly utilized strategies such as removing items that heavily cross-loaded (Pett et al, 2003; Kline, 2000). When assessing the test-retest reliability of these 54 items, we noted that 5 of these items performed noticeably worse than others. Low test-retest reliability is indicative of three potential issues, one – you may have a reliable scale but the phenomenon measured may have changed over time, two – those items may just not be reliably answered, or three – taking the test at one point may change how they answer it a second time (Streiner et al, 2015). While we cannot be 100% certain the reasons for these changes, it seems unlikely to be due to phenomenon changes (major differences in how a patient perceived their care as compassionate over a course of a single day (especially considering our question stem inquires about experiences of compassion over the past 7 days), as their relatively short period between test and retest (24 hours), and we screened for any major changes to health or healthcare situation prior to

	retesting). Although plausible, some sort of specific response bias to these items on the second testing session seems unlikely, given that the remaining 49 items remained quite stable (ICC <.70). It seems most likely to us that for whatever reason, these remaining items were simply not being answered as consistently as other items over this relatively short period of time. As such, consistent with previous literature on scale development (Hawker et al, 2008; King et al, 2007, References 28-29) we used test-retest scores to trim items. We believe that reliability is a prerequisite for later establishing scale validity. We have also now added greater detail about the various item reduction stages (Background, 3rd paragraph, 4th sentence; Phase 1 data Analysis, 2nd sentence; Strengths and Limitations, 2nd sentence) which included a Delphi process with international SMEs and patient advisors, and cognitive interviews with current patients (Reference 26).
The paper would benefit from clarifying what the "importance" ratings are about. Given the use of these ratings in the item reduction process, I would like to see the exact item.	Thank you. We have now clarified that the importance ratings were not in fact used for the item reduction process (Phase 1 Results—EFA, 3rd paragraph, 5th sentence), but were simply an indicator that our post-cut items remained very important in the eyes of participants. We apologize for the confusion on our part and appreciate the opportunity to correct this.
The information in the tables suggests a consistent degree of negative skew in the data (as well as a persistent kurtosis), presumably because people were rating their providers as very compassionate much of the time. While this is a widespread problem in research of this kind (i.e., vis-à-vis compassion in health), there is clearly a loss of discrimination at the top/positive end of the scale. Is there a reason that social desirability was not controlled and can the authors comment on the possibility of using a more "extreme" anchor (e.g., always) to fix this problem?	We appreciate the reviewer's identification of negatively skewed responses to the SCQ. While we concede there to be some ceiling effects within our measure, a portion of this may reflect true variation, simply reflecting positive experiences of compassion amongst our participants. While we certainly acknowledge that responses were negatively skewed, we note that the mode of these responses was only 4 (out of 5) for 13 of the 15 items. We agree that loss of top-end discrimination may be unfortunate, but we believe that the SCQ does a better job of this than existing measures. We also acknowledge (and appreciate the suggestion) that changing the response stems of a measure to frequency could change the response distribution, however we would point to the recently published measure on

	compassion by Roberts et al. (2019) in JAMA Network Open. This measure utilized frequency response stems (1[never] to 4[always]) but had overwhelmingly high ceiling effects (77-85% of items were responded to with “always”). For the SCQ, only 29-50% of items were responded on the top end of the response scale. Only further testing with alternative stems would confirm or disconfirm whether such changes would positively benefit the SCQ, but we remain somewhat skeptical that such a change would make a substantive difference in response distributions. We also recognize issues such as social desirability or acquiescence bias may exist within self-reported measures. To control for this, the protocol was administered by an RA who was not a member of their healthcare team and patients were specifically informed that their responses would not be shared with members of their healthcare team (we have added a sentence clarifying this (Methods-Phase 1, 1st paragraph, 5th sentence)). It was for this reason that we also recommend that the SCQ be administered by someone other than the patients’ HCP (Strengths and Limitations: 1st paragraph, 9th sentence). Unfortunately, a significant body of literature suggests that validity scales (e.g., social desirability scales, sometimes used to control for such effects) do not usually provide a useful means for detecting response distortions. A couple of important reasons for the failure of validity scales stem from social desirability scales being contaminated with substantive variance (Kozma & Stones, 1987; McCrae & Costa, 1983), and the base rate of biased responding may be relatively low (Piedmont et al, 2000). Controlling for such constructs such as social desirability, may bias results in ways that currently are unclear. We point to Piedmont, McCrae, Riemann, and Angleitner (2000) for two studies, review, and commentary supporting our position on the invalidity of validity scales. Finally, in regard to using a more extreme anchor like ‘always’, we note that we refer those readers who wish to learn more about the details of this

	decision making, to our previous publications reporting these results and provide further references substantiating these methodological decisions (Discussion, 2nd paragraph, 2nd sentence).
Data were not was (multiple instances)	Thank you, we have addressed this throughout.
Please clarify in what ways this measure represents a “gold standard”.	As noted above we have removed this overstatement on our part.
Reviewer #2	**Authors’ comments**
This is a well thought out paper that describes the need, processes, and implications of a patient-reported scale for compassion. I really enjoyed reading this article and think that The Sinclair Compassion Questionnaire addresses the need for a valid and reliable measure of compassion that has been developed with and for patients, adding to the literature in this much-needed area.	Thank you.
While the scale has been thoroughly tested using advanced psychometrics, I would suggest changing the wording slightly on page 4 line 10, and again on page 23 line 38, to this scale "could become the Gold Standard" compassion measure for patients, as other patient measures exist in the literature.	As noted above in our responses to Reviewer #1, it seems that our enthusiasm about our results resulted in some overstatements including the claim that the SCQ is a gold standard which has been removed entirely.
Editorial Comments	**Authors’ comments**
- Please revise the title of your manuscript to include the research question, study design and setting. This is the preferred format of the journal.	We have amended the title of our manuscript to “Development and Validation of a Patient-Reported Measure of Compassion in Healthcare: The Sinclair Compassion Questionnaire (SCQ) in order to align with the preferred format of the journal and other measure studies published in BMJ (i.e. https://bmjopen.bmj.com/content/9/1/e023558 **https://bmjopen.bmj.com/content/8/1/e018880)**

- Please provide a copy of the questionnaire that you have sought to validate as a supplementary file	The copyright for the questionnaire is held by UTI Partnership, the University of Calgary's innovation transfer office. In order to protect the fidelity of the measure, to monitor its usage, its accessibility is managed through a centralized website (compassionmeasure.com—this information is provided to readers in the acknowledgment section of the manuscript and the "Instruction Manual, Scoring Guidelines and Versions of the SCQ" section) where readers can access it freely once it has been accepted for publication. We have provided a copy the questionnaire for the purposes of review only.

VERSION 2 – REVIEW

REVIEWER	Consedine, Nathan University of Auckland, Faculty of Medical and Health Sciences
REVIEW RETURNED	31-Mar-2021

GENERAL COMMENTS	As previously, the submitted report presents data from a pair of instrument development studies, describing the development of a measure of the patient's experience of compassion. The revised submission shows improvement over the prior version and careful attention to the concerns raised in the initial round of review; the accompanying cover letter is thoughtful and detailed. The submission remains likely to be of high interest to the readers of the BMJ Open. I have a single suggestion for clarification: Specifically, the Abstract appears to retain some of the imprecise framing of the underlying (latent) construct that is not helpful. As mentioned in the prior review, my belief is that the construct being measured should consistently be described as the "patient experience of compassion" rather than "compassion per se. Congratulations on a really nice contribution.

REVIEWER	Durkin, Mark Department of Psychology, University of Bolton
REVIEW RETURNED	20-Apr-2021

GENERAL COMMENTS	Thank you for making the suggested edits.

VERSION 2 – AUTHOR RESPONSE

Reviewer #1	**Authors' comments**
As previously, the submitted report presents data from a pair of instrument development studies, describing the development of a	Thank you for your attention to detail. We have modified the manuscript abstract accordingly. Please note that in order to incorporate these

measure of the patient's experience of compassion. The revised submission shows improvement over the prior version and careful attention to the concerns raised in the initial round of review; the accompanying cover letter is thoughtful and detailed. The submission remains likely to be of high interest to the readers of the BMJ Open. I have a single suggestion for clarification: Specifically, the Abstract appears to retain some of the imprecise framing of the underlying (latent) construct that is not helpful. As mentioned in the prior review, my belief is that the construct being measured should consistently be described as the "patient experience of compassion" rather than "compassion per se.	changes and remain under the 300 abstract word limit, we did some minor wordsmithing to make the abstract more parsimonious. As per your suggestion, in 3 instances under the Results and Conclusions sections of the abstract (pages 2 and 3 of the manuscript), we incorporated the term 'experiences of' compassion, in keeping with the consistency of the remainder of the manuscript.