PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Study protocol for the validation of a new patient-reported outcome measure (PROM) of listening effort in cochlear implantation: the Listening Effort Questionnaire-Cochlear Implant (LEQ-CI)
AUTHORS	Hughes, Sarah; Rapport, Frances; Watkins, Alan; Boisvert, Isabelle; McMahon, Catherine; Hutchings, Hayley

VERSION 1 - REVIEW

REVIEWER	Ruth Litovsky University of Wisconsin – Madison USA
REVIEW RETURNED	25-Feb-2019

GENERAL COMMENTS	Study protocol for the validation of a new patient-reported outcome measure (PROM) of listening effort in cochlear implantation: The Listening Effort Questionnaire-Cochlear Implant (LEQ-CI) Authors: Sarah E. Hughes, Frances Rapport, Alan Watkins, Isabelle Boisvert, Catherine M. McMahon, Hayley A. Hutchings 1. Is the research question or study objective clearly defined? Not necessarily. The Introduction starts off as if the article will be a research paper, rather than a protocol for validation of an outcome measure. The introduction should be rewritten so that it is better suited for this paper. In addition, while the introduction does an adequate job of reviewing the literature and stating the aims of the study, the clarity and conciseness could be much improved. Several sentences are repetitious and need to be broken down into shorter, more specific thoughts so that it is easy for the reader to follow. □ What constitutes a psychometrically sound, effective PROM should be made clear since this will be the focus of the paper □ Could be useful to provide examples of existing hearing-related PROMs to make clearer what gap exists re: listening effort o Specific comments for introduction and aims/objectives □ Line 113- The “complex nature of listening effort construct” needs to be expanded upon since this appears to be one of the main motivations for an LE PROM. □ There should be a citation for fMRI studies of listening effort in order to be consistent with citations for pupillometry and EEG. • Why were the Zekveld and Winn studies chosen for pupillometry citations? They are neither the first nor the most recent. □ Line 116- Expanding upon the unviability of pupillometry and other measures in the audiology clinic may help strengthen the argument for the need for a PROM.

	□ Line 117- the paper states in the beginning of the introduction that there are no clinical tools to reliably evaluate listening effort and its impact on the listening activities of everyday people" but then goes on to say that there are several PROMS that are "considered to measure perceived listening effort" but just not in a manner consistent with current theoretical frameworks. The introduction may flow better if this thought was moved up towards line 104. Perhaps mention some of the PROMS that already exist to measure listening effort and the limitations of those specific PROMs. That may help to reduce the confusion. □ Line 122- It is unclear what "theoretical frameworks" the paper is referring to. Perhaps expand upon this. □ Line 137 – the paper states it will "undertake an initial assessment of the LEQ-CI's psychometric properties" by applying CTT but based on the methods, it seems that RA was the initial assessment. Perhaps remove the phrase "initial assessment" or move it to the first bullet point where RA is mentioned. 2. Is the abstract accurate, balanced and complete? Similar to comments above, there is too much emphasis on the literature without actually saying enough about the proposed study and its purpose or initial steps. 3. Is the study design appropriate to answer the research question? YES 4. Are the methods described sufficiently to allow the study to be repeated? • The methods are clear and concise, and the authors do a good job of explaining the rationale for all of the analyses they will complete. It does not seem necessary to have both inclusion and exclusion criteria included since reporting one gives information about the other. The section on discriminant validity needs to be expanded upon, as it is not clear why a moderate positive correlation is acceptable and proves that the LEQ-CI is able to discriminate between fatigue and LE or quality of life and LE. Another inconsistency is in Phase 1: CTT-Assessing internal consistency reliability. The authors states that CTT can only be used to measure the total test score validity (rather than item-level measurement), however, when discussing internal consistency reliability, they discuss inter-relatedness among items. The discussion of item specific measures in this section seems to contradict what was said early about the limitations of the CTT. • Clarifying the distinction between Rasch Measurement Theory and CTT and why both will be used would be useful • Another discrepancy occurs in the abstract, where the authors state that they will use four comparator PROMS in stage two, however, in the methods of phase 2 they only discuss two comparator PROMS (FAS and NCIQ). (line 350). • It would be very helpful to see an example of the LEQ-CI as well as the scales that the author proposes to use for each item. Overall, the methods are very well organized and are broken down in a way that is intuitive and easy to digest. While I am not an expert in validating PROMS, the author appears to have thought deeply about the proposed method and recognizes that it is an iterative process. o Specific comments for methods section:

	• Line 207 is an incomplete thought. • Lines 213-215- unclear how these are different. Is the distinction between “fits” and “describes”? • Line 225 – consider stating what the subscales are. • Line 297- consider providing a cut-off correlation value below which you would consider an item not to correlate with the overall scale and therefore be a candidate for deletion • line 351 - How weak of a correlation is considered to demonstrate evidence of discriminant validity? For both the NCIQ and FAS, where do the hypothesized correlation values come from? 5. Are research ethics (e.g. participant consent, ethics approval) addressed appropriately? Yes 6. Are the outcomes clearly defined? yes 7. If statistics are used are they appropriate and described fully? yes 8. Are the references up-to-date and appropriate? For the most part. See above for question about pupillometry citations. 9. Do the results address the research question or objective? Perhaps here or elsewhere, suggest that the authors refer to figure 2 from Hughes et al. (EAR & HEARING, VOL. 39, NO. 5, 922–934). 10. Are they presented clearly? yes 11. Are the discussion and conclusions justified by the results yes 12. Are the study limitations discussed adequately? yes 13. Is the supplementary reporting complete (e.g. trial registration; funding details; CONSORT, STROBE or PRISMA checklist)? Yes 14. To the best of your knowledge is the paper free from concerns over publication ethics (e.g. plagiarism, redundant publication, undeclared conflicts of interest)? Yes 15. Is the standard of written English acceptable for publication? yes

REVIEWER	Jani Johnson University of Memphis, United States
REVIEW RETURNED	04-Mar-2019

GENERAL COMMENTS	The field of Audiology has a notable lack of self-report measures developed with rigorous attention to psychometrics and well-established validity. I was impressed with the authors' thoughtful approach to development and validation of this questionnaire, a measure of self-reported listening effort for cochlear implant users. Such a questionnaire is needed, and no competing questionnaires are currently in standard use as far as I am aware. Rationale and methods for this study were clear, appropriate, and well-presented. I believe that this protocol will benefit the field by providing a standard for future questionnaire validation. I have no concerns about this paper and recommend acceptance for publication.

REVIEWER	Dr Amy Halls University of Surrey
REVIEW RETURNED	18-Mar-2019

GENERAL COMMENTS	I think this protocol is clearly written, with a well set out methodology and exploration and justification of processes and analysis. I have one question regarding recruitment: are the CI centres recruiting a set amount of participants each, or is recruitment starting at the same time with centres sending out as many pack as necessary until the study has 250/100 participants recruited? I think a sentence or two could be added/edited here for greater clarity.

VERSION 1 – AUTHOR RESPONSE

Reviewer Comments	Authors' Response	Line/Page Numbers (Tracked changes version)
Abstract		
2. Is the abstract accurate, balanced and complete? Similar to comments above, there is too much emphasis on the literature without actually saying enough about the proposed study and its purpose or initial steps.	We have revised the abstract to focus more on the study methods and less on the background literature. We consulted other protocols of PROM validation published in BMJ Open to ensure consistency of reporting. The revised abstract reads as follows:	Lines 44-84

	Introduction: Listening effort may be defined as the cognitive resources needed to understand an auditory message. A sustained requirement for listening effort is known to have a negative impact on individuals' sense of social connectedness, well-being, and quality of life. A number of hearing-specific patient-reported outcome measures (PROMs) exist currently; however, none adequately assess listening effort as it is experienced in the listening situations of everyday life. The Listening Effort Questionnaire – Cochlear Implant (LEQ-CI) is a new, disease-specific PROM designed to assess perceived listening effort as experienced by adult CI patients. It is the aim of this study to conduct the first psychometric evaluation of the LEQ-CI's measurement properties. Methods and analysis: This study is a phased, prospective, multi-site validation study in a UK population of adults with severe-profound SNHL who meet local candidacy criteria for CI. In Phase 1, 250 CI patients from four National Health Service (NHS) CI centres will self-complete a paper version of the LEQ-CI. Factor analysis will establish unidimensionality and Rasch analysis will evaluate item fit, differential item functioning (DIF), response scale ordering, targeting of persons and items, and reliability. Classical Test Theory methods will assess acceptability/data completeness, scaling assumptions, targeting, and internal consistency reliability. Phase 1 results will inform refinements to the LEQ-CI. In Phase 2, a new sample of adult CI patients (n = 100) will self-complete the refined LEQ-CI, the Speech, Spatial and Qualities of Hearing Scale (SSQ), the Nijmegen Cochlear Implant Questionnaire (NCIQ) and the Fatigue Assessment Scale (FAS) to assess construct validity. Ethics and dissemination: This study was approved by the Abertawe Bro Morgannwg University (ABMU) Health Board/Swansea University Joint Study	

	Review Committee (JSRC) and the Newcastle and North Tyneside 2 Research Ethics Committee (REC), Ref: 18/NE/0320. Dissemination will be in high-quality journals, conference presentations, and SEH's doctoral dissertation.	
Introduction/Abstract		
1. Is the research question or study objective clearly defined? Not necessarily. The Introduction starts off as if the article will be a research paper, rather than a protocol for validation of an outcome measure. The introduction should be rewritten so that it is better suited for this paper. In addition, while the introduction does an adequate job of reviewing the literature and stating the aims of the study, the clarity and conciseness could be much improved. Several sentences are repetitious and need to be broken down into shorter, more specific thoughts so that it is easy for the reader to follow.	Thank you. We have substantially revised the introduction in response to your suggestions. It now focuses on the measure of listening effort in the clinic and the role of PROMs more specifically. We have been critical in our appraisal of the text with the aim of enhancing the clarity of the manuscript.	Lines 103 - 158
What constitutes a psychometrically sound, effective PROM should be made clear since this will be the focus of the paper	To address this comment, we have added the following paragraph: "To have confidence that a PROM is providing meaningful information, psychometric evaluation of its measurement properties must be undertaken to satisfy rigorous criteria.[1,2] This includes assessment of an instrument's validity (i.e., does the instrument measure the construct it purports to measure), its reliability (i.e., the degree to which measurement is free from error) and its responsiveness (i.e., the ability of an outcome measure to detect change over time in the construct to be measured).[3] There are several measurement properties that require assessment (see Table 2) and each property needs its own type of study to assess it. The process of psychometric	Lines 160 – 168

	validation is iterative and represents an accumulation of evidence over time from multiple studies.[4]" We would have liked to include a table describing the various measurement properties however word limits preclude this. In lieu of a table we have included relevant references for the COSMIN guidance which explains in detail what constitutes a psychometrically robust PROM and states the criteria for the conduct of validation studies that are of high methodological quality,	Line 171
Could be useful to provide examples of existing hearing-related PROMs to make clearer what gap exists re: listening effort	This is an excellent idea and we have added a summary description of the results of a systematic review of PROMs of listening effort. Specifically, we have added the following: "Several hearing-specific PROMs have been developed that include items considered to measure listening effort. A systematic review by the authors identified two PROMs that measured listening effort and cognitive effort in listening respectively.[5,6] Several PROMs assessing listening effort at either the item or subscale level (e.g., SSQ, (A)PHAB, CPHI) were also identified.[7–10] Overall, the review findings found limited evidence of these PROMs' psychometric measurement properties. The SSQ was identified as the current best candidate for use as a listening effort PROM based on the extent and quality of its validation when assessed against the COSMIN criteria.[11] However, one drawback of the SSQ as a measure of listening effort is a high response burden with only 6% of its items measuring listening effort. Notably, all of the PROMs identified in this systematic review were developed prior to publication of the theoretical frameworks and treatises that inform current conceptualisations of	Line 139 - 150

	listening effort including the role of motivation on effort expenditure.[12–15]"	
Line 113- The "complex nature of listening effort construct" needs to be expanded upon since this appears to be one of the main motivations for an LE PROM.	Thank you for your comment. We have considered your comment fully and reached a consensus that a full discussion regarding the complexity of listening effort is beyond the scope of this paper. We have added the following: "There is a growing body of research to suggest that listening effort is a multidimensional construct and that these different measures may evaluate different aspects of this phenomenon.[16–20] Using factor analysis, Alhanbali et al. have shown that hearing level, SNR, dual-task paradigms, pupilometry and EEG (alpha power during speech recognition and retention) and self-reported effort tap into different underlying dimensions of listening effort.[17] Reflecting on this work, it may be argued that PROMs, as a measure of self-reported effort, have the potential to assess a dimension of listening effort that is not captured by current behavioural and physiological measures."	Lines 131 - 138
There should be a citation for fMRI studies of listening effort in order to be consistent with citations for pupillometry and EEG.	Thank you. As suggested we have substantially revised the introduction to focus on self-reported listening effort and PROMs. The discussion of objective measures has been removed from the manuscript.	
Why were the Zekveld and Winn studies chosen for pupillometry citations? They are neither the first nor the most recent.	This has now been removed from the manuscript	
Line 116- Expanding upon the unviability of pupillometry and other measures in the audiology clinic may	We have added a paragraph to discuss the complementary nature of self-report	Lines 129 - 138

help strengthen the argument for the need for a PROM.	measures and other measures of listening effort. "PROMs offer a complementary method to current behavioural (e.g., dual task paradigms) and physiological measures (e.g., pupilometry, fMRI, electroencephalography) of listening effort. There is a growing body of research to suggest that listening effort is a multidimensional construct and that these different measures may evaluate different aspects of this phenomenon.[16–20] Using factor analysis, Alhanbali et al. have shown that hearing level, SNR, dual-task paradigms, pupilometry and EEG (i.e., alpha power during speech recognition and retention) and self-reported effort tap into different underlying dimensions of listening effort.[17] Reflecting on this work, it may be argued that PROMs, as a measure of self-reported effort, have the potential to assess a dimension of listening effort that is not captured by current behavioural and physiological measures."	
Line 117- the paper states in the beginning of the introduction that there are no clinical tools to reliably evaluate listening effort and its impact on the listening activities of everyday people" but then goes on to say that there are several PROMS that are "considered to measure perceived listening effort" but just not in a manner consistent with current theoretical frameworks. The introduction may flow better if this thought was moved up towards line 104. Perhaps mention some of the PROMS that already exist to measure listening effort and the limitations of those specific PROMs. That may help to reduce the confusion.	We have revised the introduction to include more information on existing PROMs. The following has been added: "Several hearing-specific PROMs have been developed that include items considered to measure listening effort. A systematic review by the authors identified two PROMs that measured listening effort and cognitive effort in listening respectively.[5,6] Several PROMs assessing listening effort at either the item or subscale level (e.g., SSQ, (A)PHAB, CPHI) were also identified.[7–10] Overall, the review findings found limited evidence of these PROMs' psychometric measurement properties. The SSQ was identified as the current best candidate for use as a listening effort PROM based on the extent and quality of its validation when assessed against the	Lines 139 - 158

	COSMIN criteria.[11] However, one drawback of the SSQ as a measure of listening effort is a high response burden with only 6% of its items measuring listening effort. Notably, all of the PROMs identified in this systematic review were developed prior to publication of the theoretical frameworks and treatises that inform current conceptualisations of listening effort including the role of motivation on effort expenditure.[12–15] Lack of congruence between these instruments and current frameworks is a limitation of the content validity of existing PROMs. It is unlikely these instruments capture fully the conceptualisation of listening effort as presented in these recently published models. As such, there is growing support in the literature for a new PROM that comprehensively measures self-reported listening effort in hearing loss as it is conceptualised currently.[14,17] To address this situation, the Listening Effort Questionnaire – Cochlear Implant (LEQ-CI) has been developed. The LEQ-CI is a new hearing-specific PROM measuring perceived listening effort in adults who receive cochlear implants."	
Line 122- It is unclear what "theoretical frameworks" the paper is referring to. Perhaps expand upon this.	We note your suggestion and politely suggest that a full discussion of these theoretical frameworks is beyond the scope of what is required for an introduction to a protocol of a PROM validation study. We have included key references and noted the role of motivation which is a key iteration on the listening effort construct. We have added the following: "Notably, all of the PROMs identified in this systematic review were developed prior to publication of the theoretical frameworks and treatises that inform current conceptualisations of listening effort including the role of motivation on effort expenditure.[12–15]"	Lines 148 - 150

Line 137 – the paper states it will "undertake an initial assessment of the LEQ-CI's psychometric properties" by applying CTT but based on the methods, it seems that RA was the initial assessment. Perhaps remove the phrase "initial assessment" or move it to the first bullet point where RA is mentioned.	Thank you for your observation and yes, RA is the first assessment undertaken. To improve clarity, we have removed "initial" and revised the text so the manuscript now reads as follows: "The aim of this study is to conduct the first psychometric validation of the LEQ-CI in accordance with the internationally recognised COSMIN guidelines.[3,21]" We have removed the line "To undertake an initial assessment of the LEQ-CI's psychometric properties..." and replaced it with the following: • To assess acceptability, scaling assumptions, targeting, and reliability of the LEQ-CI using CTT methods. We note that validation is an on-going process; therefore, the use of "initial" was originally intended to mark this study as the first study undertaken to validate the LEQ-CI.	Line 170-171 Line 183
Methods		
The methods are clear and concise, and the authors do a good job of explaining the rationale for all of the analyses they will complete. It does not seem necessary to have both inclusion and exclusion criteria included since reporting one gives information about the other.	We have considered your suggestion and have amended the Table 2 in part. For clarity, we have retained "pre-lingual hearing loss" as an exclusion criterion as clarification was sought by reviewers on this point during the scientific review process.	Line 248-249
The section on discriminant validity needs to be expanded upon, as it is not clear why a moderate positive correlation is acceptable and proves	The COSMIN Initiative (www.cosmin.nl) has produced internationally recognised guidelines for the development and selection of outcome measurement	

that the LEQ-CI is able to discriminate between fatigue and LE or quality of life and LE.	instruments. The internationally-recognised COSMIN guidance [2,21,22] formed the basis for the LEQ-CI validation study protocol. COSMIN recommends a hypothesis testing approach for construct validation. Hypotheses should be determined a priori based on the literature and experience of the study team. Hypotheses should be "about expected relationships between the PROM under review and ... comparator instruments" and should specify "the expected direction (positive or negative) and magnitude (absolute or relative) of the correlations". Guidance for specifying correlation values is described in Table 8, p. 41 of the COSMIN manual [22]. There are few studies exploring explicitly the relationship between listening effort, QoL, and fatigue. Drawing from the work of Pichora-Fuller [23], Pichora-Fuller et al. [24], Alhanbali et al [19], Hughes et al. [25] and Holman et al. (in press), the study team considered these constructs to be inter-related. BY way of example, the study team refer to the work of Alhanbali et al. [19]. They used an unvalidated questionnaire – the Effort Assessment Scale (EAS) comprised of the questions measuring listening effort extracted from the SSQ and the FAS to explore relationships between effort and fatigue. Their results showed a lower correlation of 0.30 between these measures. According to COSMIN, correlations between instruments considered to measure related yet dissimilar constructs are hypothesised to be lower and in the range of 0.30 – 0.50. This guidance, complemented by the literature as discussed in the preceding paragraph, informed the process of hypothesis generation to evidence the LEQ-CI's construct validity.	

	We have made the following revisions: "COSMIN guidance specifies that construct validity may be assessed by testing a priori hypotheses based on the literature and the experience of the study team.[11] Hypotheses are generated by the study team and founded on the assumption that the LEQ-CI validly measures the target construct (i.e., listening effort). These state the relationship between the instrument and other measures, as well as the expected differences between the scores attained by different sub-groups of the target population." "As the LEQ-CI and the SSQ are measuring the same construct, we hypothesise that a strong positive correlation > 0.50 will be observed between measures as suggested by Mokkink et al.[11]"	Lines 411 - 417 Lines 426 - 428
Another inconsistency is in Phase 1: CTT-Assessing internal consistency reliability. The authors states that CTT can only be used to measure the total test score validity (rather than item-level measurement), however, when discussing internal consistency reliability, they discuss inter-relatedness among items. The discussion of item specific measures in this section seems to contradict what was said early about the limitations of the CTT.	Thank you for your comment. CTT and item-total correlations give information as regards the homogeneity of the LEQ-CI and the inter-relatedness of the items in relation to the total score. However, item-total correlations are unable to provide information on how to improve or refine items and their response scales that have been identified as potential candidates for deletion. For this reason, Rasch analysis will be applied first to attempt to rectify any problems with items and their scales. The application of CTT methods following RA is a further check of unidimensionality. This is an example of the complementary use of both RA and CTT methods.	
Clarifying the distinction between Rasch Measurement Theory and CTT and why both will be used would be useful	We have revised the manuscript substantially to address this comment. We have added the following section:	Lines 261-297

	"There are two schools of psychometric measurement theory dominate the field of PROM development.[4,26] Traditional psychometric analyses (e.g., Cronbach's alpha as a measure of internal consistency reliability) are underpinned by CTT. CTT seeks to evaluate reliability and validity of a scale and has been the dominant approach used in the development and validation of outcome measures.[27] However, modern measurement techniques such as RA are increasingly being reported alongside traditional analyses in studies of PROM development and validation (e.g., [28,29]). CTT is based on the assumption that every observed score is a function of an individual's true score and random error.[30] The assumptions underpinning CTT differ from those underpinning the Rasch model. It has been argued that CTT cannot be adequately be tested as it is based on definitions rather than assumptions which can be proven true or false. This is in contrast to modern measurement theory (i.e., RA) which can generate assumptions that can be proven true or false.[31] Whereas CTT methods focus on the total score of a measure, RA enables instrument developers to focus more specifically on the characteristics of individual items.[32] For example, RA, unlike CTT methods, can be used to establish whether an item's response scale is functioning as expected and, if not, suggest improvements. The Rasch model allows for ordering persons (i.e., patients) according to the amount of the latent target construct (i.e., listening effort) they possess and for ordering items that measure the target construct according to their difficulty.[26] This method allows non-linear (i.e., ordinal) raw data to be converted to a linear (i.e., interval) scale, which can then be evaluated through the use of parametric statistical tests.[33] By contrast, CTT methods yield measures that produce ordinal rather than interval level data. This has implications for the	

	interpretation of test scores as difference scores and changes scores are most meaningful when interval level of measurement is used.[26,31] A further limitation of CTT is that the performance of a test is dependent on the sample in which that test is assessed.[31] This renders its psychometric properties (i.e., reliability and validity) dependent on the sample rather than characteristics of the test itself. By contrast, RA produces item and test statistics that are sample independent rendering the test valid across groups. Any discrepancies between the scale data and the Rasch model requirements are indicative of anomalies in the scale as a measurement instrument. These discrepancies provide diagnostic information that serves as a basis for understanding and empirical improvement of the instrument at both item and scale-level.[34] Despite these limitations, CTT methods continue to be widely used in studies of instrument validation and are included in the COSMIN standards.[3,31] Indeed, some properties (e.g., acceptability, scaling assumptions) can only be evaluated using CTT methods.[27] For these reasons, this study will use both CTT and RA in a complementary fashion to ensure rigorous validation of the LEQ-CI at both item and scale level."	
Another discrepancy occurs in the abstract, where the authors state that they will use four comparator PROMS in stage two, however, in the methods of phase 2 they only discuss two comparator PROMS (FAS and NCIQ). (line 350).	Thank you for your comment. We are unsure why you stated only two PROMs were mentioned as the abstract mentions three – SSQ, FAS, NCIQ. However, in the abstract the error was ours, and the abstract should read three comparator PROMs + the LEQ-CI (4 PROMs in total). The abstract now reads: "...self-complete the refined LEQ-CI, the Speech, Spatial and Qualities of Hearing Scale (SSQ), the Nijmegen Cochlear	Lines 75 - 78

	Implant Questionnaire (NCIQ) and the Fatigue Assessment Scale (FAS) to assess construct validity."	
It would be very helpful to see an example of the LEQ-CI as well as the scales that the author proposes to use for each item. Overall, the methods are very well organized and are broken down in a way that is intuitive and easy to digest. While I am not an expert in validating PROMS, the author appears to have thought deeply about the proposed method and recognizes that it is an iterative process.	Thank you we have added a figure to include exemplar items from the LEQ-CI with corresponding response scales.	Figure 2
Line 207 is an incomplete thought.	With revisions to the manuscript, this line has now been removed.	
Lines 213-215- unclear how these are different. Is the distinction between "fits" and "describes"?	This line has been removed.	
Line 225 – consider stating what the subscales are.	We have given thought to this suggestion. An assessment of unidimensionality (factor analysis) is necessary in order to ascertain whether the LEQ-CI is unidimensional (1 scale) or assesses more than one construct (has subscales); therefore, we are unable to specify the subscales at this time. As such, we have taken the decision to leave this line unchanged.	
Line 297- consider providing a cut-off correlation value below which you would consider an item not to correlate with the overall scale and therefore be a candidate for deletion	Thank you for your suggestion. We have added the following sentence: "When checking homogeneity of the LEQ-CI's scales, the heuristic that items should correlate with the total score above 0.20 will be applied. Item-total correlations will	Lines 380-381

	be calculated using the Pearson product-moment correlation.[29]"	
line 351 - How weak of a correlation is considered to demonstrate evidence of discriminant validity? For both the NCIQ and FAS, where do the hypothesized correlation values come from?	Thank you for raising these questions, we have attempted to address your queries as follows: For both the NCIQ and FAS, where do the hypothesized correlation values come from? The COSMIN Initiative (www.cosmin.nl) has produced internationally recognised guidelines for the development and selection of outcome measurement instruments. The COSMIN guidance formed the basis for the LEQ-CI validation] study protocol. COSMIN recommends a hypothesis testing approach for construct validation. Hypotheses should be determined a priori based on the literature and experience of the study team. Hypotheses should be "about expected relationships between the PROM under review and ... comparator instruments" and should specify "the expected direction (positive or negative) and magnitude (absolute or relative) of the correlations". Guidance for specifying correlation values is described in Table 8, p. 41 of the COSMIN manual. [22] How weak of a correlation is considered to demonstrate evidence of discriminant validity? To our knowledge, the COSMIN guidance (Mokkink et al 2018) does not specify a minimum correlation below which is considered evidence of discriminant validity per se. Rather the authors of COSMIN state that correlations between	

	instrument scores measuring related but dissimilar constructs should be between 0.30 – 0.50 and correlations between instruments measuring unrelated constructs should be <0.30. As the relationship between listening effort and fatigue is not yet well understood but we hypothesise that the constructs of effort and fatigue are related but dissimilar, we opted to specify a low to moderate positive correlation (0.3 -0.5) as evidence of discriminant validity. We have specified the use of the COSMIN criteria and the recommended hypothesis testing approach in the text which has been expanded for clarity: "The COSMIN guidance specifies that construct validity may be assessed by testing a priori hypotheses based on the literature and the experience of the study team.[23] Hypotheses are generated by the study team and founded on the assumption that the LEQ-CI validly measures the target construct (i.e., listening effort). These state the relationship between the instrument and other measures, as well as the expected differences between the scores attained by different sub-groups of the target population.. To establish the construct validity of an instrument, Mokkink et al. recommend at least 75% of the stated hypotheses are endorsed.[23]"	Lines 411-419
Perhaps here or elsewhere, suggest that the authors refer to figure 2 from Hughes et al. (EAR & HEARING, VOL. 39, NO. 5, 922–934).	We have included the conceptual framework for the LEQ-CI as a figure (Figure 1) and referenced Hughes et al. Ear Hear, 39:5, 922–934.	Figure 1
I have one question regarding recruitment: are the CI centres recruiting a set amount of participants each, or is recruitment starting at the same time with centres sending out as	Thank you for your suggestion. We have amended the manuscript to improve clarity as follows: "In Phase 1, a cohort of 250 participants will be recruited from four National Health Service (NHS) cochlear	Lines 221 – 227

many packs as necessary until the study has 250/100 participants recruited? I think a sentence or two could be added/edited here for greater clarity.	implant centres. To minimise burden on implant centre staff and to ensure representation from different regions of the UK, each centre will send questionnaire packs to 125 cochlear implant candidates or recipients who meet the study inclusion criteria (n = 500). If necessary, additional participants will be recruited until such time as 250 completed LEQ-CI forms with no missing data are returned." In Phase 2, a new cohort of 100 participants fulfilling the same eligibility criteria will be recruited from two cochlear implant centres. Each centre will recruit 125 participants initially. If necessary, further participants will be recruited until the required sample size is achieved.	Lines 237-240