

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 of the generic, multidimensional Treatment Expectation Questionnaire (TEX-Q) through systematic literature review, expert surveys and qualitative interviews
AUTHORS	Alberts, Jannis; Löwe, Bernd; Glahn, Maja; Petrie, Keith; Laferton, Johannes; Nestoriuc, Yvonne; Shedden-Mora, Meike

VERSION 1 – REVIEW

REVIEWER	Thomas Janssens KU Leuven, Belgium
REVIEW RETURNED	20-Dec-2019

GENERAL COMMENTS	<p>The authors of this paper embarked on an ambitious and challenging research task: Their aim was to develop a comprehensive yet relatively short scale; based on extensive theory and literature review; that can be used across different treatment modalities to assess treatment expectations.</p> <p>The development of such a scale is instrumental for furthering knowledge on changes in expectations and their relevance for treatment outcomes. The questionnaire also has the potential to serve as a manipulation check for interventions that aim to change treatment expectations.</p> <p>The research paper that describing the result of this task does however suffer from a series of limitations. I therefore would like the authors to comment on the following issues:</p> <ol style="list-style-type: none">1. The paper describes the development process of the TEX-Q scale, starting from a theoretical model and systematic evaluation of existing scales, and resulting in the construction of a 35-item questionnaire. Although the authors describe several steps in this process, I feel that some aspects of this process may need more detail in order to evaluate the quality of the research.<ol style="list-style-type: none">1. There is only limited information on the generation of the conceptual structure of the tex-q. In what way did the resulting literature review extend the original theoretical model by Laferton et al. (2017). Which additional constructs were discovered and evaluated in the literature review but not retained in the model?2. There are only limited details available on the process of going from 538 potential items based on the literature review to 78 items that were assessed for content validity in unclear; both in terms of process (the open iterative discussion may need more clarification), as well as the resulting 78 items (how did these items match on the pre-defined domains?)3. The overall high approval ratings for the questions strike me as discordant with the comments in the commentary section. Is this due to item effects (small numbers of items are deemed to be performing
-------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

	<p>poorly) or due to rater effects (small number of raters having criticisms), or some other potential reason?</p> <p>4. The cognitive interviews and qualitative inquiry into patients concepts of treatment expectations may need additional detail. which questions were used during the interview? was this a structured/semi-structured interview? is the interview guide available? Which parameters were used for sample selection and was data collection continued until saturation or based on a pre-defined criterion?</p> <p>5. in the current paper, the authors do not evaluate the psychometric characteristics of the resulting questionnaire. It is currently not clear why psychometric evaluation was not used to aid item reduction. Furthermore, information about reliability and validity of the scale in different use contexts and sensitivity to change may be needed in prior to using the tex-q questionnaire in different research contexts.</p> <p>2. in the introduction, the authors conceptualise expectations as a construct that is related to probability. However, in the resulting items of the questionnaire, participants are not asked to evaluate probabilities (e.g. how likely do you think negative treatment outcomes are to occur), but only assess the magnitude of expected (positive/negative) treatment outcomes. I agree with the authors that changing all questions to a degree of magnitude of change makes the questionnaire more easy to use, but it seems a major step away from their theoretical framework.</p> <p>3. a concept that is included in the theoretical paper by Laferton et al. (2017) is strenght of expectations. this concept also relates to bayesian belief models (confidence in a belief/precision of a belief), and may help to unite the probabilistic view on expectations (probability that an event will occur) vs. the magnitude based view on expectations (expected magnitude of change).</p> <p>4. (minor) In the example question for feared harm, the anchors (degree of harm) do not seem to match the question (degree of fear)</p>
--	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

REVIEWER	Alexandra Kern Institute for Complementary and Integrative Medicine, University Hospital Zurich, Switzerland
REVIEW RETURNED	05-Jan-2020

GENERAL COMMENTS	<p>Alberts et al. used a three-step approach to develop a generic, theory-based, multidimensional measure that can be used to assess patients' expectations of medical as well as psychological treatments.</p> <p>Overall, the manuscript is very well written and the process that lead to the development of the measure is clearly/sufficiently described. Moreover, the abovementioned three step development approach represents a sound method in order to develop an assessment tool of this kind.</p> <p>However, there are two suggestions to further improve the quality of the manuscript.</p> <ul style="list-style-type: none"> - While the fact that the development of the measure is based on a theory (in this case the Integrative Model of Expectation) is very valuable, the downsides of such an a priori restriction are not sufficiently described in my opinion. I would suggest to provide a short rationale for this initial decision or at least to shortly discuss the matter in the discussion section. - Page 7 Line 35 and Figure 3: It is not clear, where the additional 33 relevant articles are coming from. What exactly was your "additional search strategy"? I would recommend clarification here.
-------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

Reviewer Name: Thomas Janssens

Institution and Country: KU Leuven, Belgium Please state any competing interests or state 'None declared': None declared

The authors of this paper embarked on an ambitious and challenging research task: Their aim was to develop a comprehensive yet relatively short scale; based on extensive theory and literature review; that can be used across different treatment modalities to assess treatment expectations.

The development of such a scale is instrumental for furthering knowledge on changes in expectations and their relevance for treatment outcomes. The questionnaire also has the potential to serve as a manipulation check for interventions that aim to change treatment expectations.

The research paper that describing the result of this task does however suffer from a series of limitations. I therefore would like the authors to comment on the following issues:

1. The paper describes the development process of the TEX-Q scale, starting from a theoretical model and systematic evaluation of existing scales, and resulting in the construction of a 35-item questionnaire. Although the authors describe several steps in this process, I feel that some aspects of this process may need more detail in order to evaluate the quality of the research.

1.1. There is only limited information on the generation of the conceptual structure of the tex-q. In what way did the resulting literature review extend the original theoretical model by Laferton et al. (2017). Which additional constructs were discovered and evaluated in the literature review but not retained in the model?

Response: In general, we found all of your following comments extremely helpful, as they pointed to missing or unclear information that is relevant for the reader to follow our development process.

Thank you very much for such a thorough and valuable feedback. Thus, we modified our manuscript accordingly and hope that we have been able to elucidate the mentioned criticisms.

Regarding your first comment, we thoroughly revised the respective section both in the methods and results to provide a more detailed description of this process.

p.5, l.20-24.: The conceptual structure of the TEX-Q was developed based on the most relevant expectation theories (22, 23) incorporated in our Integrative Model of Expectations (18). Our goal for the conceptual structure was to cover a relevant range of treatment-related expectation constructs with potential predictive value for outcome. At the same time, we aimed to include concepts that can be generically assessed.

p.7, l.3-39: The literature review generally provided additional support for the Integrative Model of Expectations (18), with all reviewed items fitting to one or more of the aspects of expectations differentiated in the model. Some of the scales reviewed, however, focused more specifically on one or more aspects of the model and therefore introduced additional, more nuanced differentiations within it. Our rationale was to capture the most potentially relevant aspects of this expectations with predictive value for outcome in the TEX-Q. Hence, we developed a 2x2x2 concept to operationalize outcome expectations in the questionnaire (Figure 2).

Firstly, we distinguished probabilistic expectations, describing realistic assumptions about what is likely to happen (e.g., expecting symptom improvement) from value-based expectations, describing more affective, less rational feelings like hopes or fears (e.g., hoping to be pain-free). Our rationale here was to capture the potentially different predictive value of these expectation constructs as

theorized in the literature (16, 18). Secondly, we distinguished expectations about beneficial outcomes (e.g., treatment success) from expectations about harmful outcomes (e.g., complications). This inclusion was based on empirical evidence from the literature pointing to these aspects being separate dimensions rather than two sides of a unidimensional structure (30). Thirdly, we distinguished expectations about direct, symptom-related treatment outcomes (e.g., benefit or side effects) from expectations about the broader impacts of the treatment (e.g., improved quality of life or reduced functioning). We thereby introduced a categorial operationalisation of the range of possible treatment outcomes described in the Integrative Model of Expectations as relevant for different treatment outcomes to secure generic applicability (18, 22). The eight terms depicted in the central cells of Figure 2 describe the resulting theorized subscales of the TEX-Q. In addition to the aforementioned scales measuring outcome expectations, we included two additional subscales. The first was process related expectations (e.g., a straight-forward procedure), based on the assumption the expectations and experiences of the treatment process will be related to treatment outcome particularly in long-lasting treatments (18, 23, 24). The second was the expected behavioural control of the treatment (e.g., being able to influence treatment success), based on the rationale these capture situation-specific correlates of generalized self-efficacy (31). In total this led to ten different theorized subscales for the TEX-Q.

We refrained from the inclusion of further nuanced views from the conceptualisation of the TEX-Q for the sake of its applicability and generic nature. We also excluded general expectation constructs from our conceptualisation, as the TEX-Q was planned to focus on expectations about medical and psychological treatments, and good measures for relevant general expectation constructs like self-efficacy or optimism already exist (32, 33). Furthermore, we had to exclude the timeline dimension of treatment expectations due to the dissimilarity of timelines in different treatments and the resulting lack of potential generic formulations possible in its operationalisation.

1.2. There are only limited details available on the process of going from 538 potential items based on the literature review to 78 items that were assessed for content validity in unclear; both in terms of process (the open iterative discussion may need more clarification), as well as the resulting 78 items (how did these items match on the pre-defined domains?)

Response: We added further information about the process, aim and results of our discussion process.

p.8, l.17-23: Based on this list, all authors took part in an iterative discussion process about the construction of the scale and its potential items. In that process, the items were further evaluated regarding their fit to our conceptual model, their applicability as generic items and our overall impression of them. Several items were reformulated to make them more generic and additional ones were constructed. With the deletion of duplicates and those substantially overlapping content wise, as well as those we consensually found to not fit our conceptual model, we selected 78 items, each clearly associated with one of our conceptual subscales. These provided the basis for the further development of the TEX-Q.

1.3. The overall high approval ratings for the questions strike me as discordant with the comments in the commentary section. Is this due to item effects (small numbers of items are deemed to be performing poorly) or due to rater effects (small number of raters having criticisms), or some other potential reason?

Response: Only approximately one third of the items was criticized by the raters, and even in the case of any specific criticism, ratings were rather positive. We clarified the number of items the comments were targeting in the results section.

p.8, l.31-32: The commentary section of the ratings contained criticism about the wording of approximately one third of the items.

4. The cognitive interviews and qualitative inquiry into patients concepts of treatment expectations may need additional detail. which questions were used during the interview? was this a structured/semi-structured interview? is the interview guide available? Which parameters were used for sample selection and was data collection continued until saturation or based on a pre-defined criterion?

Response: We fully acknowledge that – for the sake of brevity – the information we provided was rather superficial. We added more detail on our recruiting process in the methods section. In addition, we now attach our structured interview guide as Appendix B.

p.6, l.15-23: In the selection of these patients for the interviews, we aimed to maximize the diversity of conditions and treatments. Patients were interviewed by male and female researchers with prior experience with this assessment (JA, MG). Data saturation was discussed regularly and data collection was continued until we found it to be sufficient within this sample.

Based on a semi-structured interview guide (see Appendix B), the patients were asked to complete the potential TEX-Q items, some of them in different phrasings to examine the differences, while speaking out their thoughts (thinking-aloud technique (28)).

1.5. in the current paper, the authors do not evaluate the psychometric characteristics of the resulting questionnaire. It is currently not clear why psychometric evaluation was not used to aid item reduction. Furthermore, information about reliability and validity of the scale in different use contexts and sensitivity to change may be needed in prior to using the tex-q questionnaire in different research contexts.

Response: Psychometric evaluation of the TEX-Q is an integral part of the development of the TEX-Q and will thereby also be used for further item reduction, but this paper focused on the contentual development of the TEX-Q. As this was a thorough and long process, we felt that we could not meet the reporting standards of each development step sufficiently, if we had decided to also include the - quite extensive - psychometric evaluation in this paper. Psychometric results will be published elsewhere. We added a sentence in our manuscript to clarify this.

p.13, l.8-11: Further validation of the scale in different clinical settings is necessary to confirm the psychometric properties of the TEX-Q and possibly further reduce its number of items. Therefore, the psychometric evaluation with patients from four different clinical settings will be published elsewhere.

2. in the introduction, the authors conceptualise expectations as a construct that is related to probability. However, in the resulting items of the questionnaire, participants are not asked to evaluate probabilities (e.g. how likely do you think negative treatment outcomes are to occur), but only assess the magnitude of expected (positive/negative) treatment outcomes. I agree with the authors that changing all questions to a degree of magnitude of change makes the questionnaire more easy to use, but it seems a major step away from their theoretical framework.

Response: Thank you for this important remark. Please see our combined answer to the following comment no. 3.

3. a concept that is included in the theoretical paper by Laferton et al. (2017) is strenght of expectations. this concept also relates to bayesian belief models (confidence in a belief/precision of a belief), and may help to unite the probabilistic view on expectations (probability that an event will

occur) vs. the magnitude based view on expectations (expected magnitude of change).

Response: Indeed, your comment hits one of the most challenging aspects we faced in developing the items. We discussed this point thoroughly and repeatedly at several stages during the TEX-Q development process. Both the probability and the strength of expectations are relevant and potentially different aspects in patients' treatment expectations. The best way of how to capture both aspects of probability and strength of expectations would be to assess them with one item each (e.g., "How likely do you think your symptoms will improve?" and "How much improvement do you expect?").

In addition to Kirsch's theoretical considerations regarding the probabilistic nature of non-volitional expectations (Kirsch, 1985), we considered the results of the in-depth development process of a questionnaire measuring patient expectations of complementary and alternative medicine (Sherman et al., 2014). Sherman et al. conclude that phrasing the items using the term "realistically expect" best captures both aspects of strength and probability. Following this line of thinking, we tested different formulations in our cognitive interviews. As a result, we framed all probabilistic expectations with the following introductory text: "The following section is about your expectations regarding your treatment. Please try to assess them as realistically as possible."

We think that with our phrasing we have found the best compromise to capture probabilistic as well as strength-related aspects of expectations in a way that is acceptable for the patients and economically assessable. We added the following information describing this process in the results of the interview.

p.10, l.12-15: Especially aspect 4 was broadly discussed with the aim to capture both the probabilistic and strength related aspects of expectations in an easily comprehensible way, resulting in the reformulation of all remaining items into a change question format and the addition of a brief introductory text asking patients to assess their expectations "as realistically as possible".

4. (minor) In the example question for feared harm, the anchors (degree of harm) do not seem to match the question (degree of fear)

Response: We understand your reading of this item, but do not agree that it refers to one's degree of fear, but rather to the degree of feared risk. We furthermore assume that this slight ambiguity in the formulation will be clarified to the patients through the anchors.

Reviewer: 2

Reviewer Name: Alexandra Kern

Institution and Country: Institute for Complementary and Integrative Medicine, University Hospital Zurich, Switzerland Please state any competing interests or state 'None declared': None declared

Alberts et al. used a three-step approach to develop a generic, theory-based, multidimensional measure that can be used to assess patients' expectations of medical as well as psychological treatments.

Overall, the manuscript is very well written and the process that lead to the development of the measure is clearly/sufficiently described. Moreover, the abovementioned three step development approach represents a sound method in order to develop an assessment tool of this kind.

However, there are two suggestions to further improve the quality of the manuscript.

1. While the fact that the development of the measure is based on a theory (in this case the Integrative Model of Expectation) is very valuable, the downsides of such an a priori restriction are not sufficiently described in my opinion. I would suggest to provide a short rationale for this initial decision or at least to shortly discuss the matter in the discussion section.

Response: Thank you for this important comment. We added this as limitation of our theory-based development in the discussion.

p.12, l.39-42: Although the theory-based construction of the questionnaire grounded on the Integrative Model of Patient Expectations (18) provided a valuable framework for its subdimensions, it may thereby also have led to the exclusion of additional dimensions that could have emerged in a purely empirical concept development.

2. Page 7 Line 35 and Figure 3: It is not clear, where the additional 33 relevant articles are coming from. What exactly was your “additional search strategy”? I would recommend clarification here.

Response: These articles resulted from the additional critical review of treatment-specific expectation scales. Additionally, we searched for scales measuring expectations of specific treatments through the assessment of the literature list from our critical review article on patients’ expectations, as well as the literature lists of further treatment specific reviews that were mentioned in our article. We added more information to refer to our additional search strategy.

p.8, l.5-6: The additional critical review of treatment-specific expectation scales lead to the inclusion of further 33 relevant articles, resulting in a total of 9345 articles.

VERSION 2 – REVIEW

REVIEWER	Thomas Janssens KU Leuven (University of Leuven)
REVIEW RETURNED	10-Mar-2020

GENERAL COMMENTS	<p>Thank you for your responses to my comments on the original manuscript. I feel that the additional details provided by the authors greatly improve my understanding of the questionnaire development process, and also signals the rigor applied in the development.</p> <p>I have one additional/continued remark: Reviewing the (novel, back-translated) example questions, there remains an ambiguity in whether the questions ask about magnitudes of change (symptom improvement or worsening); or if the questions are about probability of the occurrence of positive/negative outcomes. It is possible that this is due to the ambiguity of the word “fear” and “risks”, which i am biased to understand as a probabilistic concepts</p> <p>Apart from the example question about feared harm, the example questions for the constructs of expected harm, feared negative impact, process related expectations, and expected behavioural control also share this ambiguity.</p> <p>To some extent, i feel this ambiguity is unavoidable if the aim is to evaluate an aggregate of magnitude as well as probability. However, there is a risk in individual differences in interpretation of the items, as this may cause items to perform poorly by introducing additional error variance, or result in correlations between items that is based on participants conceptualisation of the items. I therefore feel that a discussion about treatment expectations being related to both probabilities of outcomes as well as magnitudes of outcomes and implications of this for the assessment of expectations could be part of the discussion section.</p>
-------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

REVIEWER	Alexandra Kern Institute for Complementary and Integrative Medicine, University Hospital Zurich, Switzerland
REVIEW RETURNED	29-Mar-2020
GENERAL COMMENTS	My two suggestions have been addressed to my full satisfaction.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 1

Thomas Janssens
KU Leuven (University of Leuven)

Thank you for your responses to my comments on the original manuscript. I feel that the additional details provided by the authors greatly improve my understanding of the questionnaire development process, and also signals the rigor applied in the development.

I have one additional/continued remark:

Reviewing the (novel, back-translated) example questions, there remains an ambiguity in whether the questions ask about magnitudes of change (symptom improvement or worsening); of if the questions are about probability of the occurrence of positive/negative outcomes. It is possible that this is due to the ambiguity of the word “fear” and “risks”, which i am biased to understand as a probabilistic concepts

Apart from the example question about feared harm, the example questions for the constructs of expected harm, feared negative impact, process related expectations, and expected behavioural control also share this ambiguity.

To some extent, i feel this ambiguity is unavoidable if the aim is to evaluate an aggregate of magnitude as well as probability. However, there is a risk in individual differences in interpretation of the items, as this may cause items to perform poorly by introducing additional error variance, or result in correlations between items that is based on participants conceptualisation of the items. I therefore feel that a discussion about treatment expectations being related to both probabilities of outcomes as well as magnitudes of outcomes and implications of this for the assessment of expectations could be part of the discussion section.

Response: Thank you for your further remark on this topic. We fully acknowledge the ambiguity you pointed out and included a discussion of it in the limitations section of our manuscript, incorporating your argumentation as well as the reasoning we laid out in response to your comment 3 in our first response letter.

p.12, l.39 – p.13, l.5: Other important limitations lie on the conceptual level. Our item phrasing is ambiguous with regard to two different aspects of expectations, assessing the magnitude of an expected change as well as the expected probability of its occurrence to some extent. The relevance of this differentiation was pointed out in the Integrative Model of Expectations (18) and is further supported by recent empirical evidence (70). Based on Kirsch’s (22) theoretical considerations regarding the probabilistic nature of non-volitional expectations, and on Sherman’s (71) in-depth analysis of patients’ understanding of treatment expectations, our phrasing aimed to capture best both aspects of magnitude and probability. To prevent ambiguity, each of these aspects could have been assessed in a separate item (e.g., “How likely do you think your symptoms will improve?” and “How much improvement do you expect?”), but we refrained from it for the sake of our scale’s brevity and applicability. This may lead to differences in individual interpretations of our items and thereby lower their quality.

Reviewer: 2

Alexandra Kern

Institute for Complementary and Integrative Medicine, University Hospital Zurich, Switzerland

My two suggestions have been addressed to my full satisfaction.

Response: Thank you very much, we are glad to hear that.