Peer Review File

Article information: https://dx.doi.org/10.21037/apm-23-462

Reviewer A

Comment 1: The then-test method has been debunked because it suffers from low reliability due to recall bias, low validity such that the "information" it detects relates only rarely to recalibration response shift, and low interpretability due to the fact that researchers using it never manage to find a way to express their results other than "over-emphasis" and "under-emphasis" depending on whether the then-test score is larger or smaller than the pre-test score. Indeed, a recent scoping review by leaders in the field of response shift research EXCLUDED articles using the then-test because of its problematic reliability and validity. So it is hard to understand the rationale for the authors' systematic review, particularly because they state most of these weaknesses and then proceed to try to summarize and interpret the "findings" as though they are meaningful. Thus, at square one of the manuscript, I would have stopped reading were I a normal reader rather than a reviewer.

Reply 1: Thanks for your comments. I understand the then-test may be problematic when assessing response shift. Beyond summarizing the findings from these papers that used the then-test, we will make an effort to also report on each paper's difficulties when using the then-test. This we hope will inform the reader that the then-test is inadequate and as you mention- to not encourage its continued use (at least without the proper analysis).

Changes in the text 1: Another paragraph in the discussion was added (fourth paragraph). The methods used to assess response shift were summarized, emphasizing that most of the articles used the then-test only, therefore exposing them to bias. Moreover, it was summarized in the same paragraph how many articles mentioned recall bias in their limitations and an additional method to detect response shift was suggested as a way to help control for this bias.

Comment 2: There is no rationale given for the dates of the literature search (2005-2019). Why start so late and end several years before the manuscript's submission?

Reply 2: Thanks for your comment. I understand the dates of the search may seem arbitrary. After looking into this, we erroneously reported in the results that the latest article included was published in 2019, when in fact the included study by Ten Ham was published in 2020. After revisiting the data from the literature search, the search had a span from 2005 to the time of the search.

Changes in the text 2: The text was edited to correct the error that the latest article was published in 2019. See page 1, line 34. It was also clarified that the search was carried out from 2005 to the time of the literature search in May 2022. See page 1, line 32.

Comment 3: The fact that the then-test ONLY putatively measures recalibration response shift is in itself a weakness.

Reply 3: Thank you for pointing this out. We will make this clear.

Changes in the text 3: Please see the addition at line 264, page 6: “Another limitation is that the then-test only putatively measures recalibration. Ten articles used the then-test only which raises suggesting these particular studies missed response shift effects for reprioritization and reconceptualization.”

Comment 4: The conclusions based on the putative findings are pointless. The “presumption that response shift phenomena obfuscate treatment benefits” could be more rigorously documented using solid methods, not methods that should be ignored going forward. Indeed, publishing a systematic review of the then-test would only serve to encourage the unenlightened to continue using it.

Reply 4: Thanks for your comment. I agree that this paper may encourage some to continue using the then-test. We will make it clear that assessing response shift should be an undertaking done with rigor and sound methods- that means using methods less prone to bias and limited in scope.

Changes in the text 4: The text was modified (line 257, page 6) to indicate that other methods than the then-test should be used to assess response shift or if the then-test is used, other methods should be included as well.

Comment 5: The description in the Discussion of other methods used to detect response shift (last paragraph page 5) is embarrassing in its typos and misunderstanding of the actual methods used. It leaves out important new developments such as the application of random effects modeling with equating (companion papers by Schwartz et al, Quality of Life Research 2021) which, by the way, is feasible with small samples. It neglects to even mention the twenty-five years of research on appraisal which the senior author has been involved in.

Reply 5: Thank you for your comments. We apologize for the typos and misunderstanding of the actual methods used. In this paragraph, other modalities for measuring response shift are discussed. We realize now that this information should touch upon past research of appraisal of the then-test. In accordance with your suggestions, we think it is wise to add a paragraph summarizing the difficulties the authors had with the then-test, such as its inconsistency with results with other tests, or the limitations they listed from using the test.

Changes in the text 5:
At the end of the third paragraph of the introduction, criticisms of then then-test are listed- including how social desirability, implicit theories of change, and recall bias may explain away the findings.

In the second paragraph of the discussion on page 5, a note is made to recommend the use of other methods to assess response (or at least use an additional test) considering the weaknesses of the then-test.
In the third paragraph of the discussion on page 5, an alternative method to assessing response shift in smaller sample sizes is given as suggested in your comment. This paragraph also relists other methods used for measuring response shift (ensuring the names are of the methods are correctly written).

Comment a:
Providing a comprehensive summary of the methodological issues that have documented in favor of abandoning the use of the then-test. This summary would need to be added to the second to last paragraph on page 2 and would include all of the papers documenting recall bias (see work by Ahmed, Schwartz, Bloem, and many many others); noisy measurement using qualitative and quantitative methods (see Schwartz et al. published in Quality of Life Research, 2012;21(3):381-388),

and the general interpretability problem as described in Schwartz and Sprangers Quality of Life Research 2010; 19:455-464.

Reply a:
Thank you for the comment. We agree these issues are important. A summary of research that documents the limitations of the then-test have been addressed under comment 5.

Changes in the text a:
Please see the end of the second to last paragraph on page 2.

Comment b.
Implementing the systematic review with an eye toward highlighting the issues raised in point (a) above and noting how each of the ill-fated cancer research papers using the then-test suffered from one or more of these issues. This would be a novel take on the idea of a systematic review. It would also avoid the problem of "any attention is good," which has the risk of encouraging the continued use of this inadequate method.

Reply b:
Thank you. We understand your comments and we will restructure our position on the use of the then-test to not risk encouraging its continued use. We also addressed this issue under comment 1.

Changes in the text b:
We summarized the weaknesses of the then-test that these papers stated they suffered from. This was also addressed under comment 1. In the fourth paragraph of the discussion on page 6, the specific limitations these studies outline are summarized. An addition was added to the conclusion (line 265-67) to add caution to the use of the then-test: "While the then-test has contributed to some patient perspective of cancer patient outcomes, this study has shown that the then-test, if used, should be used with caution considering its limitations and the emergence of more advanced methods."

The following sentences have been deleted:
Lines 265-66: "Nevertheless, this study reviews the usage of the then-test, and these studies have contributed to the understanding of response shift and provided a basis for further research."
The last sentence of the manuscript text (lines 276-77): "This review has demonstrated the value that the then-test has added to the understanding of outcomes in cancer patients from the patient's perspective" has been deleted.

Reviewer B

Comment 1: Title: „impact of then-test analysis on quality of life outcomes". This formulation suggests that the then-test would have a direct impact on HRQOL, which is not the case. The then-test has no impact on patients HRQOL, but just helps to detect a response-shift. Response-shift as well has no effect on HRQOL outcomes themselves, but only on the reports/ratings/results/scores provided by patients on a scale/PROM.

Reply 1: Thank you for your comment, this is good advice. We will change the title to show that the then-test helps detect response shift and does not directly impact HRQOL. We have changed the title

to “A Systematic Review of the Use of the Then-Test for Evaluating Health-Related Quality of Life in Cancer Patients"

Changes in the text 1: Please see page 1, lines 1-2

Comment 2: Abstract: The methods section says „studies were included if they 1) used the then-test to measure response shift“. If this is an inclusion criterium, authors can delete the sentence „All measures used the then-test to detect response shift“.

Reply 2: Thank you for the comment, we will delete this sentence and say “studies were included if they 1) used the then-test”

Changes in the text 2: Please see page 1, Line 32

Comment 3: Introduction: The introduction reads very well and provides a good background for this review. The only suggestion I have is to add the word previous in line 80: „to retrospectively give a judgement of their previous HRQoL based on their current perspective“.

Reply 3: Thank you for your suggestion! We had added the word previous to line 80.

Changes in the text 3: Please see page 2, line 80

Comment 4: Search strategy: The PRISMA guidelines are reporting guidelines and no guidelines on conducting systematic reviews. Please, reformulate the first sentence accordingly.

Reply 4: Thank you for clarifying, we have changed the sentence to say “The results of the systematic review are reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.”

Changes in the text 4: Please see page 3, lines 97-98

Comment 5: If the PRISMA reporting checklist will be published as supplement, refer to it in the text.

Reply 5: Thank you. This will be added.

Changes in the text 5: A line was added on line 98, on paragraph 1 in the methods: “A PRISMA checklist is published as a supplementary file.”

Comment 6: Data collection: Please, add more information on the „significant differences/results“ that were extracted. Which effect sizes, scores, and indicators of uncertainty / significance were extracted?

Reply 6: Thank you for pointing this out. “Significant differences/results” should be specified.

Changes in the text 6: “Significant differences/results” has been replaced with “response shift effect sizes with standard deviation” on page 3, line 116.

Comment 6-a: In the results section authors refer to mean differences, but also to “response shift scores” which I haven’t heard of before. As these results are at the core of this review, please, explain its calculation and interpretation in the methods section.

Reply 6-a: Response shift effect sizes or scores using the then-test is the difference between the then-test score and the pre-test score. If the score is negative, the respondent scored their then-test as less than the pre-test. If the score is positive, the respondent scored their then-test as higher than the pre-test. Lines 82-84 in the introduction currently reads: “Therefore, the difference between the then-test and pre-test indicates the degree of response shift and the difference between the then-test and post-test provides evidence of true change in QoL (2).”

Changes in the text 6-a:
In the methods at line 117, a sentence that reads as the following is added: Response shift effect size using the then-test is defined as the difference between the then-test and the pre-test scores.

RESULTS:

Comment 7: Evidence of response shift: First, it took me a moment to understand that the first paragraph focuses on prostate cancer, the second on breast cancer, and the third on cancer patients in general. To guide the reader, authors could just start each paragraph with a brief sentence for orientation, e.g., “XX studies focused on prostate cancer”, “XX studies investigated response shift in breast cancer”.

Reply 7: Thank you for your suggestion on improving clarity of the results. We have added introductory sentences for the results paragraphs.

Changes in the text 7: Please see page 4:
Line 157: “Response shift in prostate cancer patients was measured in a study done by Ten Ham et al(n = 3,161) using the then-test (16)”
Line 174: “Two studies investigated response shift effects in breast cancer patients.”
Line 195: “The following studies evaluated various types of cancer patients.”

Comment 8. Some acronyms for PROM titles, which were introduced in the section ‘measurement tools’ are re-introduced again in the results section.

Reply 8: Thank you for noticing this error, we have changed to just the abbreviations after introducing each PROM.

Changes in the text 8: Please see page 4, line 159-160

Comment 9. Line 162-163: “The PCI resulted in a greater response shift than the SF-36”. Is it possible to compare response shifts observed on different measures directly to each other? Or is this result depending on the scoring algorithms and scale of each measure? E.g., if one measure has a range from 0-28 and another measure ranges between 0-100, then probably the mean response shift on the 0-100 scale is going to be larger compared to the 0-28 scale. Also, two measures with the same range (e.g., 0-100) can differ in their granularity if they have a different number of items and response options per item. Thus, the smallest possible response shifts on both measures might be different.

Reply 9: Thank you for this comment. We agree that these two scales are not comparable being two different scales with different granularities. Therefore, comparing magnitudes is less appropriate.

Changes in the text 9: To address this issue, we rephrased the sentence so the results are stated, but not compared to each other. Lines 162-163 reads now: “The PCI response shift score was 7.4 (range, 4.3 to 14.7) and the mean SF-36 response shift score was 15.1 (range, 4.0 to 30.2) (25)”

Comment 9-a. If a direct comparison is possible, the formulation “the PCI resulted in a greater response shift” is not 100% correct. Clarify, e.g., by reformulating “A greater response shift was detected on the PCI compared to the SF-36”.

Reply 9-a: Thank you for your recommendation on rewording this comparison. We decided against directly comparing the two.

Changes in the text 9-a: Refer to changes for comment 9.

Comment 10. Line 171-173: Is this sentence related to the PRO-ratings on patients’ current HRQOL at each time point or are these results somehow taking the detected response shift into account?

Reply 10: The sentence “Results showed that IPSS and SPI significantly decreased over 6 months.” indicates the change without response shift: from pre-test to post-test. The key finding in this paper was that the retrospective then-test assessment was higher than this original change. It is not clear how it is currently written how response shift ties in to this and we will clarify the statement.

Changes in the text 10: Lines 171 to 173 is now put into better context and reads: “Results showed that IPSS and SPI significantly decreased over 6 months from pre-test to post-test. The then-test scores were consistently higher than the pre-test scores with statistical significance for both these scales, indicating response shift possibly occurred (23).”

Comment 11. Line 187: “The recalibration effect was statistically significant”. At first, I wondered how we know that this is a recalibration effect instead of reprioritization or reconceptualization. It is only mentioned in the discussion that the then-test is meant to quantify recalibration specifically. Authors could provide this information already in the introduction.

Reply 11: Thank you for this suggestion. We have included this in the introduction: “The then-test specifically quantifies recalibration.”

Changes in the text 11: Please see page 1, line 77

Comment 12. Line 197-198: “Emotional functioning significantly improved using the then-test approach.” Comparable to the formulation in the title, this is not correct. Instead, emotional functioning is just rated better in retrospect, it is not improved in itself by using the then-test.

Reply 12: Thank you for pointing this out, we have altered the wording of this line so that this conclusion is accurate. It now says “Emotional functioning significantly rated better in retrospect improved using the then-test approach 14.45 (95% CI: 11.22-17.67), while fatigue, pain, and global QoL significantly deteriorated.”

Changes in the text 12: Please see page 5, lines 197-198

Comment 13. Line 202-205: On some scales, there seems to be both, an increase and decrease of response shift. How is that possible?

Reply 13: Thank you for this inquiry. The Ousmen study evaluated magnitude of response shift effect, it's direction (whether respondents over- or underestimated their pretest scores) and how the absolute value of the response shift increased or decreased. They noted a trend which was most apparent in patients that rated worse quality of life over time, that the absolute value of the response shift increased over time for these patients. We will clarify this result in the text.

Changes in the text 13: We adjusted the text at lines 202-205: "Ousmen et al. found that the magnitude of the response shift effect increased over time in patients whose QoL deteriorated and decreased in patients who reported improvement of QoL (26). These changes were most notable in patients whose QoL deteriorated showing that the magnitude of response shift increased between 3 months and 6 months in 13/15 dimensions of QLQ-C30 questionnaire and 4/7 dimensions of QLQ-BR23 questionnaires, thereby providing evidence that response shift may have a greater impact if patients report declining QoL (26)."

Comment 14. Line 209 "found that the Cantril then-test ratings were lower than the pre-test ratings". For readers, it is not clear if higher scores indicate better or worse functioning or higher or lower symptom burden. Probably, this is different for different measurement tools referred to in this review. Authors should check whether all results are clear in this regard.

Reply 14: Thank you for this comment. We will clarify how the Cantril ladder is scored.

Changes in the text 14: Line 209 of the results has an addition explaining the Cantril ladder score.

Comment 14-a. Line 269-273 in the discussion: QOL scores for functioning were higher at T3 compared to T2, but symptoms scales higher at T2 compared to T3. In this case, for example, this is due to the direction of the scales. While higher scores on functioning scales refer to better functioning, higher scores on symptom scales refer to higher symptom burden. Thus, in both cases patients rated their HRQOL better at T3 compared to T2.

Reply 14-a: Thank you for pointing this out. We agree that overall for this, patients rated their HRQOL as better at the retrospective assessment performed at 6 months looking back at the 3 month mark compared to the standard score at the 3 month mark. We wanted to make a point that timing of retrospective assessments and multiple timepoints are important to consider when assessing for response shift. To address this, we think it is best to delete this portion written in the conclusion and add to the discussion to make our point.

Changes in the text 14-a:
Lines 267-275 have been deleted. A section has been added in the new 4th paragraph of the discussion. It adds to the discussion of recall bias being a limitation to the then-test indicating the difficulty with choosing appropriate time points and how studies may adjust for multiple time points when assessing for response shift.
The addition is: "Timing of follow-up assessments may influence results. Assessments conducted too soon to baseline patients are still adjusting to treatment and its effects, whereas too long in the future

and issues with memory may arise. A study included in our review by Hamidou et al. assessed how TTD of QoL changed with or without accounting for then-test results (7). Several other studies in our review used multiple timepoints to assess response shift. Sébille et al. (2021) have suggested that response shift detection methods such as SEM, mixed models, Rasch Measurement Theory, and Item Response Theory are accommodating to multiple timepoints (30)."

Comment 15. Tables 2 and 3: Why are the results only reported in details for these two measures, the EORTC QLQ-C30 and the SF-36? And what exactly is meant with "response shift indicators"? Which scores are displayed in these tables (mean differences, …)?

Reply 15: Thank you for your comment, this is important to clarify. The EORTC QLQ-C30 and the SF-36 were the most common measures used across all the studies so they are the chosen to be reported in separate tables. This was done to ensure comparability between scores. "Response shift indicators" represent which functioning scale the response shift is detected on (for example, physical functioning vs role functioning). We understand this may be vague and will rephrase.

Changes in the text 15:
The table headings for tables 2 and 3 now read: "mean differences of response shift values" instead (Response shift values being difference between then-test and pre-test values). The table has also been adjusted to specify that the brackets reflect standard deviation.

DISCUSSION

Comment 16. Line 225 "Our study has demonstrated the importance that response shift has" sounds as if this was an original research article actually demonstrating response shift effects. However, this review rather provided an overview of the existing literature using the then-test to investigate response shift effects.

Reply 16: Thank you for pointing this out. We have changed the first sentence of the discussion to ensure we encapsulate the fact that our study is a review of existing literature. The first line now reads "Our study provides an overview of existing literature using the then-test to investigate response shift effects."

Changes in the text 16: Please see page 5, line 225-226

Comment 17. Overall, a large part of the discussion refers to other methods to analyse response shift and to why they are considered more versatile than the then-test. Thus, the discussion reads mostly like a limitations section. While it is important to highlight this limitation and to refer to other methods, authors missed some opportunities to actually discuss their own results and findings and to interpret their implications for the field. Some aspects are already mentioned briefly. I suggest to clarify and expand the discussion of these:

Comment 17-a. Line 226-227: "Patients in some studies recalled their pretreatment HRQoL as better … other reported them as worse" – Add numbers of how many studies found better/worse retrospective ratings. Is there a tendency that retrospective ratings are mostly better or worse? (consider the direction of the scales as mentioned in comments before)

Reply 17-a: Thank you for your comment. There is no tendency for retrospective ratings to be mostly better or worse. It is clear there are a lot of factors that influence the direction of the changes.

Changes in the text 17-a: Added to the first paragraph of the discussion is the following: “Overall, 9/16 studies showed response shift impacting QoL in both directions. For studies that measured overall health using a specific scale such as “global QoL”, “general cancer distress”, or “general state of physical health”, the results were varied. Three studies scored overall health as higher retrospectively compared to the baseline score, one study showed no change, and six studies scored overall health as lower retrospectively.” This takes into consideration the direction of the scales.

Comment 17-b. Line 231-232: “response shift … may have implications on the efficacy of their treatment”. As mentioned in some comments before, this formulation is not correct. Response shift has no effect on treatment efficacy, but just on PRO results. This complicates the interpretation of these results and thus can lead to over- or underestimating the efficacy of a treatment. Take your time and words to explain this relationship between response shift and treatment evaluation.

Reply 17-b: Thank you for this suggestion. We have altered the discussion to clarify applications of response shift on PRO results. The sentence now reads “It is important to consider response shift in a patient’s disease course because it may have implications on patient reported outcomes (PRO’s) (20). This concept can be applied when considering how a treatment may alter aspects of a patient's life, and thus allow more personalized treatment.”

Changes in the text 17-b: Please see page 5, lines 231-232

Comment 17-c. Line 246: “continues to be used frequently”. Is that statement supported by the publication dates of the included studies?

Reply 17-c: Thank you for your comment. We will specify that it has been commonly used in the cancer patient population.

Changes in the text 17-c: The beginning of paragraph 3 of the discussion now reads:
Despite its limitations, the then-test has been commonly used to assess response shift in cancer patients. A systematic review in 2019 of response shift in studies assessing quality of life in cancer patients showed 21/35 studies used the then-test (30).

Comment 17-d. Line 267: “Results of this review indicate that the timing of retrospective assessments may have an impact on how patients evaluate certain quality of life indicators”. This is an important finding, on which authors could elaborate more. At the moment, the subsequent sentences (lines 268-273) read like a section from the results again and refer only to one specific study. In the discussion, authors should try to combine results from all studies included in the review and draw conclusions from their combination. Did other studies find the same?

Reply 17-d: Thank you for your comment, this is an important point. The timing of retrospective assessment matters because of recall bias. Some retrospective assessments were done days apart, while others were completed months or years apart. Typically patients were assessed at 3 and 6 months post treatment. However, after 6 months it may become more difficult for patients to accurately recall their HRQOL.

Changes in the text 17-d:
In the first paragraph of the discussion, I show that results of the then-test are mixed and response shift is occurring in multiple directions. I chose to summarize and compare results that assessed changes in general quality of life as it was the most comparable scale among different studies.
The following was added to first paragraph of discussion: “For studies that measured overall health using a specific scale such as “global QoL”, “general cancer distress”, or “general state of physical health”, the results were varied. Three studies scored overall health as higher retrospectively compared to the baseline score, one study showed no change, and six studies scored overall health as lower retrospectively.”
I believe that results are more comparable when results from different studies are compared using the same follow-up time point. To make the results more comparable, we adjusted table 1’s results to include the results at 3 months follow-up for the studies by Ousmen et al. and Anota et al.
Also, in the fifth paragraph of the discussion this is now written: “Timing of follow-up assessments may influence results. Assessments conducted too soon to baseline, patients are still adjusting to treatment and its effects, whereas too long in the future and issues with memory may arise.”

Comment 17-e. Beyond that, authors could highlight in the discussion, that only one of all included studies was conducted in pediatric oncology.

Reply 17-e: Thank you for this suggestion, we can include this information. The following sentence was added “In addition, only one of all included studies was conducted in pediatric oncology. Thus, determining how different patient populations may experience response shift is another aspect that should be investigated”

Changes in the text 17-e: Please see page 6, line 257

Comment 18. Line 250-251: “found that applying SEM to evaluate response shift … had a recalibration effect for social functioning”. As in some comments before, the reformulation is not correct. Rather: “applying SEM… detected a recalibration effect”.

Reply 18: Thank you for your comment, we have changed this line to make it more accurate. The line now reads “For example, Friedrich et al. found that applying SEM to evaluate response shift in breast cancer patients detected a recalibration effect for social functioning (22).”

Changes in the text 18: Please see page 6, lines 250-251

Comment 19. In general, authors should provide a good justification for their focus on the then-test. Already in the introduction of the abstract, the shortcomings of this test are highlighted. So, why did the review focus on this method in particular?

Reply 19: This review was done in order to outline existing literature on the then-test which evidently is still a commonly used modality for assessing response shift in cancer patients. We think reviewing its use is important to better inform readers and authors of its uses, but more importantly its limitations concerning its bias and other improved methods. We want to emphasize that caution should be used when employing the then-test to assess response shift.

Changes in the text 19:

In response to reviewer A, a review of the limitations these studies documented with the then-test is summarized. This is at the end of the third paragraph of the introduction.
On line 257, page 6, recommending other methods than the then-test should be used to assess response shift or if the then-test is used, other methods should be included as well.
In addition, from lines 276-77, “This review has demonstrated the value that the then-test has added to the understanding of outcomes in cancer patients from the patient's perspective”, is now deleted.