Author's response to reviews

Title: Developing and testing an instrument to measure the presence of conditions for successful implementation of quality improvement collaboratives

Authors:

Michel LA Duckers (m.duckers@nivel.nl)
Cordula Wagner (c.wagner@nivel.nl)
Peter P Groenewegen (p.groenewegen@nivel.nl)

Version: 2 Date: 29 April 2008

Author's response to reviews: see over
Dear editors,

On the 26th of February we received your email with the comments of the reviewers on our manuscript. Based on their remarks and findings we decided to completely revise the manuscript. This means that we now focus on measuring the organization of QIC teams and the way they are supported instead of the general conditions for successful QIC implementation. Furthermore, we ran the analyses in a larger sample.

In this second manuscript many changes were made compared to the previous one. In our opinion this has led to a substantial improvement of the manuscript now carrying a new title: ‘Developing and testing an instrument to measure team organization and team support within quality improvement collaboratives’. We would like to thank both reviewers for their careful study of the manuscript and will reply to the different issues raised by them. Our request to you is to consider the adapted manuscript for publication.

Kind regards,
Also on behalf of Cordula Wagner and Peter Groenewegen

Michel Dückers

Reviewer 1

1. When creating a new instrument it is easier to remove items rather than add new ones, which would require a new sample to test them. Data should be collected on at least five but fewer than 10 items for each dimension theoretically postulated to define the construct. This recommendation is based on Guilford's rule that at least three items are needed to define a factor, balanced with what is known about reliability estimation. There are numerous formulas that demonstrate that the more items on a measure, all other things being equal, the more reliable the measure. Two of the factors identified in this instrument are comprised of only two items.

Reaction: In the new study Guilford’s rule is met; the components are comprised of at least four items. Data is collected on five items.

2. Although the instrument was demonstrated to be reliable and evidence of content validity was provided, the only evidence of construct validity is inter-item and inter-factor correlations. There is no statistical evidence that the measure is positively related to other constructs that should be positively related (e.g., conditions for successful implementation). How do we actually know that this instrument will achieve its main purpose, which is to assist in adopting quality improvements? To assess this, collection of additional data (measurable outcomes) is required which can be associated with the surveys and subsequently analysed.

Reaction: The main use of the altered instrument is to measure three dimensions of QIC implementation: team organization and internal and external support. Although literature suggests that the three dimensions in the current study contribute to successful implementation of QIC projects (e.g. Leggatt & Dwyer 2003; Greenhalgh et al. 2005; Mills & Weeks 2004; Øvretveit et al. 2002), an actual assessment of the degree of success is indeed very valuable, however, we do not yet have all the outcome data for all the teams in our data set.

3. The factor structure is questionable and confidence in the instrument is low. One of the main factors for this uncertainty is the small sample size. Comfrey and Lee (1992) suggest that
the adequacy of sample size might be evaluated very roughly on the following scale: 50 very poor; 100 poor; 200 fair; 300 good; 500 very good; 1000 or more excellent. Gorsuch (1983) recommends a minimum subject to item ratio of at least 5:1 in exploratory factor analysis (EFA), but notes that higher ratios are generally better. There is a widely-cited rule of thumb from Nunnally (1978) that the subject to item ratio for EFA should be at least 10:1. In the case of this study it is 101:30 or 3.4:1, which is well below the recommended standard. Bryant and Yarnold suggest there should be at least 10 cases for each item being used, the subjects-to-variables ratio should be no lower than 5 and the total number of cases used in EFA should not be less than 200 (Bryant and Yarnold, 1995).

Reaction: We added a third group of project leaders to the dataset on which the analyses are based. Also we reduced the number of items from 30 to 15. The new subject to item ratio is now 165:15 or 11:1. This is an important improvement compared to the analyses in the first manuscript. With regards to the sample size: we admit that 163 is not as much as we would have liked to have at our disposal, nevertheless there are two reasons why in our case it should not have to be a problem. According to MacCallum et al. samples between 100 and 200 can be good enough provided that communalities are higher than .5 and there are relatively few factors each with only a small number of indicator variables. The lowest communality in our study is .52 and we found three factors with four to five items.

4. Another problem is the diversity -combined with small sub samples- of the professional groups used for testing the instrument. Although all subjects were regarded as project leaders, their different professional backgrounds imply different perspectives and opinions. Although the authors acknowledge this as a potential limitation, it still remains a problem. Furthermore, the authors appear to easily overcome the problem by stating that despite these differences, it is likely that the instrument is suitable for. On what arguments is this assertion based?

Reaction: The project leaders in our sample have different occupations. Roughly these can be divided in three groups: advisors/administrative or quality staff, management/department head or medical professional. Our assumption that the instrument is suitable for other evaluation studies is based on the finding that these respondents, despite their different backgrounds and functions, have the same tasks in the programme and they could very well distinguish the three dimensions.

5. The title of the paper suggests that an instrument is developed and tested that measures the presence of conditions for successful implementation of Quality Improvement Collaboratives. Indeed, a questionnaire is developed which depicts whether several processes, support, etc. are present. But it does not show that the scale scores achieved affect success and to what extent they can predict it. In fact, the authors leave this for future research. Therefore, isn’t it necessary to show that the literature presented on pages 4-5 indeed proves their success empirically and, in addition, whether evidence exists that the same conditions can be transferred to another setting internationally?

Reaction: See the comment on issue 2. The theoretical and empirical literature suggests that team organization and internal and external support influence the success of QICs. To be able to test this assertion for Dutch QICs we developed a questionnaire to measure these influences.

6. On page 8, (line 9) it is mentioned that 101 questionnaires were returned. On the same page (next paragraph) it is also said that 3 respondents had filled in less than half of the total number of items and were excluded from further analysis. Yet on page 9 (Results Section) again 101 leaders are mentioned, whilst adding up 25 physicians, 20 advisors, 15 unit heads, 6 nurses, 3 pharmacists and 31 other positions gives us a total of 100. Which figure is the correct one (101, 98 or 100)?

Reaction: In the revised study the correct figure is 168. After excluding 3 questionnaires 165 questionnaires remain.
7. On page 11 the authors suggest they removed items 11 and 29 from further analysis when it is obvious from Tables 3 and 4 that items 11 and 30 have been in fact removed. A more close examination of items should take place since other errors seem to exist. For instance, in Table 5 item No 28 is mentioned twice (within factor 1 and 7).

Reaction: Indeed, it appears we have not been careful enough in the first manuscript. In the second manuscript only one item is removed. We double checked the index numbers.

8. Table 6 could be simplified, since the range of correlations between items and other factors seems to give redundant information (already given in Table 5). Also, some of the footnotes of the Table give the same information as can be found in the main text.

Reaction: The range of correlations is corrected for overlap and therefore slightly different from the information in the earlier table. An alternative format is chosen to present the multi trait/multi item matrix.

9. The authors state that no items were excluded based on the percentage of missing responses, although a missing value of about 15% was observed. Why? Isn’t this a high percentage that might imply problems with the clarity of the question?

Reaction: The highest percentage of missing responses in the new data set is 4.8%.

10. The scale of item 23 is said to have been inverted (Table 3, page 21). What does this mean? Typically, Likert scales allow one to interchange positively and negatively worded items in order to minimise any positive biases. Has this been done in only one question here?

Reaction: Several of the items on executive leadership are no longer part of the instrument as presented in the revised manuscript. Item 23 is one of the items we removed.

11. Why hasn’t test-retest reliability been assessed? This should be mentioned in the limitations section.

Reaction: Good point. The main reason we did not perform the same survey with the same respondents at a different moment has to do with agreements made between the funding organization, the hospitals, the programme makers and us as independent evaluation researchers to minimize the questionnaire burden for hospital staff. This is mentioned in the limitations.

Reviewer 2

1. The authors extracted seven components from a 30 item questionnaire. With such a high number of extracted components in a small item pool follows a typical problem: Two of the seven components become scales that contain only two items. A psychometric scale with two items is controversial. Although the seven factors are distinguishable in the principal component analysis, since the PCA was employed to create scales, scale aspects need to be considered as well. The authors should at least address this problem in the text or state how they will deal with this issue. The possibilities are to accept two item scales, to add similar items to create a longer scale, to discard the two factors, or to critically review the decision to extract seven factors from a 30 item pool. Although there was internal consistency in this sample, there are reduced chances that such a short scale will achieve acceptable internal consistency in another sample.

Reaction: This problem is solved, also see the reaction on issues 1 and 3 from the first reviewer.
2. It is not clear how consistent the cross-loading issue was approached. Item 11 and 29 were removed but why were items 27 and 28 left in the questionnaire when they had substantial secondary loadings on the first component?

Reaction: Not longer an issue. The scope of the article is changed. Only three factors remains.

3. The implied superiority of an answer mode that contains a middle category instead of a "risky" 4-point scale is not warranted in my opinion. Using a neutral answer is perfectly legitimate but it should not be overlooked that it comes with its own set of problems. It is no secret that a middle category threatens the underlying interval assumption of a rating scale (3 is more than 1 and less than the score 5) since a middle score can either indicate a middle position on an interval scale or indicate that the rating scale was not used (so being outside the dimension that is assessed because the respondent couldn’t answer)

Reaction: we agree with the second reviewer. The remark on 4-point scales is removed.

4. The method section reports some results—the number of missing items and one measure of dispersion characterizing the answer mode. Since this is a legitimate issue when evaluating the quality of an item this may be better addressed in the result section (i.e. items that produce a lot of missing answers are not useful).

Reaction: See the reaction on issue 9 of the first reviewer. In the result section it is now emphasized that no items were excluded based on the percentage of missing responses.

5. The measures associated with table 3 are missing in the result section, i.e. the analysis section mentions a measure of dispersion but the result section paragraph is missing.

Reaction: Table 2 shows the distribution of the answers given by the respondents, together with mean, median and standard deviation.

6. The factor analytic write-up would benefit from adding the scree plot. The decision to extract seven factors from a questionnaire that only contained 30 items could be considered controversial. Although the authors report that they looked at the scree plot, I think it would be easier for the reader to follow that decision were the scree plot included in the result section. Especially since the first 10 or so eigenvalues were not reported so that the reader would have another indication of why seven components needed to be extracted.

7. There are different criteria in what to look for in a scree plot inspection. The authors mention that they looked at where the plot abruptly levels out. This may confuse Cattell’s approach of identifying scree, which does not necessarily show an abrupt change of level and the more pragmatic approach of looking for a visible gap in the scree plot. I would suggest the authors clarify this and/or report the scree plot.

Reaction to issue 6 and 7: A scree plot is added. Together with the Kaiser-Guttman rule this should give enough information to support our decision to extract three components.

8. I think it would be more appropriate not to call the analysis a factor analysis in the paper but a component analysis since the employed PCA does not find underlying factors but the contained components that are entailed in the sample; this would apply to all text and tables.

Reaction: Factor analysis is replaced by principal component analysis.

9. It seems the method section does not entirely match the results section. Some of the planned analyses do not seem to have been reported.
10. The order of mentioned analyses in the method and result section do not really match which may make it harder for the reader to follow the paper.

Reaction to issue 9 and 10: We double checked this and believe to have made some improvements by rearranging several parts of the text.

11. The factor analysis approach takes up most of the result section with classic reliability and item analytic considerations being mainly reduced to the convoluted table 6. I personally think it would be worth looking at other reliability measures than Cronbachâ’s alpha and to report them clearly. The method section suggests that corrected item total correlations will be presented but these are not that easy to find in the result section. It could be argued that table 5 overwhelms the reader with information that mainly is not critical (since the loadings for all items on all components are already reported in table 4) and the important information (item-totals (?) for each scale) drowns in the number of results.

12. It may be better to find an alternative wording for ‘corrected for overlap’ in table 5 (it is not entirely clear what that means and how it was achieved).

Reaction: The second table now also includes scales if item deleted. Moreover, we are not exactly sure if it is necessary to replace ‘corrected for overlap’. We provided additional information to explain the procedure (a correlation corrected for overlap is the correlation of an item with the sum of the other items in the same scale; the bias of correlating an item with itself is thus removed).

13. Initially the authors assumed a five factor model. In case the authors tried to replicate these five factors rather than being guided entirely by the exploratory analysis, this should be reported. It may be worth looking into setting the number of extracted components to five and to evaluate the emerging components conceptually. Especially if the authors insist on the seven component structure it may make a stronger case were other models investigated apart from following entirely the exploratory and data-driven nature of the procedure.

Reaction: absolutely relevant, but not longer an issue.

14. This paper addresses the factorial validity of the designed questionnaire. A very important point, however, I think it is worth pointing out that this is not the only criteria a measurement tool has to fulfil. The items were derived from previous research and by consulting experts what do mainly practitioners think what are characteristics for successful implementations of quality improvement interventions. Quality improvement research is only at the beginning in building up a constructive evidence base. There are not many certainties yet, we do not know a lot about what kind of quality improvement intervention or components works. Studies that systematically investigate the variables that predict successful short term and long term implementations are not as numerous yet as that they could have meaningfully informed the initial item pool. To me it seems that the authors demand a leap of faith from the reader that the initial item pool consisted of relevant items and covered all relevant aspects. From the current study we only know that there is empirical evidence that practitioners are able to distinguish the mentioned factors. Whether these are in fact related to successful implementation of quality improvement interventions needs to be seen. The authors address this point as further research, however, it could be made clearer that so far we donâ’t know whether the questionnaire really measures what it was designed to measure.

Reaction: See the reaction to issues 2 and 5 of the first reviewer.

15. Although the decision to use a Varimax rotation was understandable and was justified in the text, it might be worth mentioning the conceptual problem here: are the factors really meant to
be independent conceptually. It might be worth adding a sentence that the focus was on identifying components that can be distinguished empirically.

Reaction: After having given the matter some extra thought we came to the conclusion that team organization and internal and external support not necessarily are independent. On the contrary, it is more logical that team organization is higher when support is higher. We therefore applied an oblique promax rotation. According to Gerbing and Hamilton (1996) both the Varimax and Promax rotation will yield similar results. Both pattern and structure matrices are presented.

References: