Author's response to reviews

**Title:** Designing questionnaires: Healthcare survey to compare two different response scales

**Authors:**

Salome Dell-Kuster (salome.dell-kuster@usb.ch)  
Esteban Sanjuan (esteban.sanjuan@usb.ch)  
Atanas Todorov (atanas.todorov@usb.ch)  
Heidemarie Weber (heidemarie.weber@usb.ch)  
Michael Heberer (michael.heberer@usb.ch)  
Rachel Rosenthal (rachel.rosenthal@usb.ch)

**Version:** 5  
**Date:** 21 July 2014

Author's response to reviews: see over
Dear Dr. Ziegenfuss, Dear reviewers, Dear editorial board,

Thank you for your further effort to review and comment upon our manuscript. All issues that you raised have been developed in the following point-by-point reply. We highlighted all changes to the manuscript in yellow.

We appreciate that the quality of the manuscript has further improved due to the changes and additions suggested by the reviewer. We therefore would be grateful if the manuscript now meets the criteria for publication in BMC Medical Research Methodology.

Kind regards,

Salome Dell-Kuster MD MSc

Point-by-Point reply

Answers to the reviewers
We thank the reviewer for their valuable comments. Below we address these comments in a point-by-point reply. The revised part of the manuscript is given in italics for each of the comments.

Reviewer 1

Major compulsory revisions:
1. Previous comments on compulsory revision
I find the previous comments for compulsory revision are fully reflected in the current version. For instance, the definition of ceiling and floor effect are still absent and I do not find statistical significance tests with the results. For instance, on pg. 10, the authors state, “Patients who completely filled out all questions (Webappendix: Additional file 3) were representative in baseline characteristics for those who sent back the questionnaire (table 1).” Where is the statistical test that supports this statement? Not providing statistical tests seriously limit readers’ understanding of the extent to which the findings of this study apply.

Reply:
Definition of floor and ceiling effect
In the background section, we added a definition of the floor and ceiling effect to the revised version of the manuscript as follows:
A ceiling effect occurs when a measure possesses a distinct upper limit for potential responses and a large number of patients score at or near this limit (the opposite of a floor effect). A high ceiling effect renders discrimination at the high-end difficult and reduces the possibility of measuring further improvement.

Statistical significance test for baseline characteristics
Many thanks for this important feedback. We agree with the reviewer that stating that the patients who completely filled out all questions were “representative” of those who returned them may suggest that actually some statistical tests have been carried out.

As a matter of fact, we a priori chose to refrain from carrying out statistical tests to compare baseline characteristics for the following reasons:
1) Conducting significance tests comparing tables of baseline characteristics would actually violate the STROBE guidelines for reporting observational data (http://www.strobe-statement.org/; Vandenbroucke et al., Epidemiology 2007; von Elm et al., Lancet 2007). The STROBE guidelines state: “In studies that compare groups, the descriptive characteristics and numbers should be given by group, (...) and significance tests should be avoided in descriptive tables.“ In the statistical analysis section of our manuscript we outlined that we adhere to these guidelines.
2) Due to the large dataset, even very small differences are likely to become statistically significant, although this is by far not clinically relevant.
3) Additionally, carrying out hypothesis tests on every single baseline characteristic will lead to a multiple testing situation increasing the chance of committing an alpha error (type I error, i.e. concluding that there is a difference, when in fact there is no difference) (Neuhäuser, Fandam Clin Pharmacol 2006).

By simply comparing the distribution of the baseline characteristics of those patients who completely filled out all questions and those patients who returned the questionnaire, most of the numbers are the same or show only minimal differences. Even without performing any statistical test, we think that it is fair to say that the patients who completely filled out all questions are similar to those returning the questionnaire.
In order to address the reviewer’s concern, i.e. to render it clearer that we did not perform any test, we rephrased the sentences as follows:
In the abstract: Baseline characteristics of the responders and non-responders were similar.

In the results: Patients who completely filled out all questions (Webappendix: Additional file 3) showed similar baseline characteristics as those who sent back the questionnaire (table 1).

2. Presentation of the response order
With the full questionnaires, I see that two difference between NS and LS questionnaires. One is the number of the response points that the authors indicated as an experimental factor; and the other is the order of response options. For instance, responses to a question, “How would you rate the quality of treatment received?” were ordered from poor to excellent in the NS questionnaire but from excellent to poor in the LS version. It is well documented that order of response options matter as respondents are more likely to choose
those presented first (e.g., Carp, 1974; Chan, 1991; Krosnick and Alwin, 1987). This is clearly shown in the domains of question to physicians, question to nurses, and respect and dignity reported in Table 2. Because the differences between two versions are product of both the number of the response points and the response order, the findings between NS and LS are the product of the two. In other words, the response order differences may confound what the authors planned to examine. This should be further examined or pointed out as a potential confounder.

Reply:
We agree with the reviewer that we are faced not only with two different kinds of response scales but also a different polarity of the response scale. Therefore, we have already stated in the discussion section (limitations [p. 15] and findings in relation to other studies [p. 16]) that the different polarity of the response scale may partly explain the divergent answers resulting in a low correlation coefficient and a higher ceiling effect of the LS, where the positive answers were on the left-hand side. We agree with the reviewer that the reversed answering scale may be the reason for the slightly higher ceiling effect of the LS (as stated in the discussion on page 16) and for the few observations with contradictory responses on the two scales (for all domains less than 1%). However, most of the participants (over 80% of the patients) rated the 5 questions on the numeric scale with at least an 8 and almost 90% (87-95% depending on the question) with at least a 7, even though on the numeric scale the negative answers were presented first. Assuming a strong influence of the reversed numeric answering scale, the response distribution on the numeric scale should have been much more symmetrical. As the reviewer points out, the questions with the lowest percentages in the positive answers were the domains of question to physicians and nurses. On the other hand, in the domain of whether treated with respect and dignity, the pattern of responses on both scales resembles more the other two domains about quality and return to the same hospital (see figure 4 and 5). The reason why we did not observe such a strong influence by the polarity might be that we used a numeric scale with higher numbers corresponding to higher quality of care. This scaling is rather intuitive since already the Swiss school system functions with higher number corresponding to higher marks. This ordered (reversed) response scale is a distinct difference to the studies by Krosnick et Alwin and by Chan. Krosnick et Alwin asked for ranking qualities of children from a unsorted list of characteristics. Chan investigated the effect of reversing an adjectival response scale. We suspect the effect of polarity would have been more pronounced with the adjectival scale presented with the worst answers first and the numeric scale with the best answers first, but obviously we cannot prove this with our data. Since this is a pragmatic evaluation of two existing questionnaires, we were unable to change the questionnaire-scales and could not investigate the influence of these two features separately.

In the revised version of the manuscript, we describe the difference in polarity already in the method section as follows:

**NS:** The response scale is displayed on an 11-point numeric scale (NS) with anchors at both ends, presenting the negative answers first.

**LS:** The response scale is displayed on an adjectival scale, with, according to the question, three or four labelled response categories, presenting the positive answers first.
We further expanded the discussion adding that the differences between the answers on both response scales could be confounded by the fact that the numeric response scale was reversed by adding to the discussion of the revised manuscript:

*People tend to use the first satisfactory response option in a presented questionnaire, as shown in studies which directly compared a reversed to a not-reversed answering scale (Krosnick et Alwin, Public Opin Quart 1987; Chan, Educ Psychol Meas 1991). However, had the reversed answering scale of both questionnaires a strong influence in our setting, we would expect the response distribution of the numeric scale (with negative answers on the left-hand side) to be more symmetrical.*

**Minor essential revisions:**

1. Does “lower confidence interval” on pg. 11 mean the lower bound of confidence interval?

**Reply:**

Thank you, that is what we meant. The revised manuscript now reads:

*Cronbach’s alpha for all five questions was 0.77 on the LS with a lower bound of the confidence interval of 0.755 as opposed to 0.89 on the NS with a lower bound of the confidence interval of 0.886.*

2. What do “within-centre” and “between-centre” mean in the last sentence of the abstract?

**Reply:**

We called the comparison of the same hospital over several years the ‘longitudinal within-centre comparison of hospital care’. On the other hand, we named the comparison of hospital care between different hospitals at the same point in time a ‘cross-sectional between-centre comparison of hospital care’.

To render it clearer, the sentence now read as follows:

*Otherwise, a longitudinal within-hospital or a cross-sectional between-hospital comparison of patient care is questionable.*