Reviewer's report

Title: Adaptation and validation of the Treatment Burden Questionnaire (TBQ) in English using an internet platform

Version: 1 Date: 1 May 2014

Reviewer: Timothy Hofer

Reviewer's report:

2014-06-BMC-Adaptation and validation of the Treatment Burden Questionnaire (TBQ) in English using an internet platform

I was asked by the editorial staff to provide an additional, rapid turn-around review focusing on the statistical methods as a complement to other reviews already received. I do comment on other elements that I noted in passing while reading the rest of the paper but assume that other reviewers have paid more attention to the other parts of the paper.

Section-Abstract

Abstract: Context: you might say "Here, we aimed to evaluate the validity and reliability of an English version of the TBQ, a scale originally developed in French."

Abstract: Conclusions: how about something more like: "We demonstrate that the English TBQ is a reliable instrument in this population and provide evidence supporting the construct validity for its use to assess the burden of treatment for patients with multiple chronic conditions in English-speaking countries."

The reason for "in this population" : The reliability of an instrument is conditional on the population in which it is tested (given that reliability is equivalent to the between person variance divided by the total variance). In a more homogeneous population the between person variance is smaller and the reliability is less. In a more heterogeneous population it is the reverse. Reliability is of course also conditional on the instrument as a better instrument will have less residual error and thus more reliability (as the residual is part of the total variance in the denominator). But its dependence on the population has to be kept in mind, and gives you an incentive to describe the population and to remind users that in a different population it might have a different reliability.

Section-Background

end of 2nd para: How about "developed" in stead of "evaluated" in "The instrument was first evaluated in a sample of patients with multiple chronic conditions recruited in hospitals and general-practitioner clinics in France."

Section-2.3 Pretesting
2.3 Para 3 "Concerning missing items, 15 patients (7.5%) found that patient–caregiver relationships were insufficiently covered in the original items. Other suggestions were either specific to a particular condition, related to the burden of disease or were already covered in the existing items. Thus, we added a new item for testing: "

"Concerning missing items" is a little confusing and I initially thought you were talking about missing data. Maybe be less telegraphic and say "Patients were asked whether there were any important elements of treatment burden that they saw as missing from the questions currently on the questionnaire"

Section-2.4

The series of steps undertaken to evaluate the measurement properties of the instrument are clearly described and seem appropriate. I am satisfied with the statistical methods and reporting. Doing test-retest reliability is admirable and certainly represents going "above-and-beyond" what almost scale developers usually do.

Following the lines of the comment above about the importance of defining the population in which the reliability is characterized, it would be good to spend some time providing any available information and a discussion about how representative the surveyed population from the PatientsLikeMe network is of the ultimate population in which the instrument is intended to be used. There is a sentence or two in the discussion that describes how well the PatientsLikeMe resemble the general population but is this the intended population for using the instrument? I might comment a bit more about this. National healthcare surveys might provide some information on characteristics of the population seeking healthcare to compare to the surveyed population. I would at least provide some caveat about how the reliability is conditional on the heterogeneity observed in this population.

2.4 Para 5. "Construct validity was tested by confirming 4 pre-supposed hypotheses." would change "pre-supposed" to "pre-specified" maybe

It wasn't clear to me how items for which the respondent replies "Does not apply" are scored. Can you stick that in somewhere? (now the scoring seems to be covered primarily in the introduction unless I missed where else it is specified)

Section Conclusion

end: "The present study advances the evidence of the validity of the TBQ; however, further work is needed to evaluate its performance in clinical settings" This is perhaps more vague than you need to be. What do you mean by "performance"? Really the next step is to look at its longitudinal responsiveness, does it change when a persons burden of treatment would seem like is should change? Or if optimistic, you could jump right in and look at whether an intervention designed to minimize burden leads to a lower score in the experimental group than the control group following the intervention."
Overall I think this is very nicely done. Probably my most important comment is about describing the population assessed as compared to the population(s) in which the instrument is intended be used.

Although I don't recommend it for this paper at this point, when you do this complete a reliability analysis you do have an option of doing a combined analysis where you look at the several key variance components (or facets of measurement) simultaneously, so you describe variance across items, occasion (the test-retest), country and maybe even disease.

Then you can define the reliability for a variety of measurement settings of interest fixing different facets of measurement- e.g.:

1/ 15 items, at one occasion with random samples of disease and country
2/ 15 items, at one occasion for patients with a single disease and in a single country.
3/ 7 items, at one occasion, with random samples across disease and country.

etc

Given that all those reliabilities will be different it also further emphasizes how important it is to define the measurement conditions or procedures under which you intend to use the instrument and how much the measurement properties depend on the measurement conditions & procedure.

If interested see:


Also some venerable but still relevant examples in the medical literature

Evans WJ, Cayten CG, Green PA. Determining the generalizability of rating scales in clinical settings. Med Care 1981 Dec;19(12):1211–20. (not quite as clearly written as it could be)

and


Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.
Declaration of competing interests:

I declare that I have no competing interests