Author's response to reviews

Title: On line clinical reasoning assessment with the script concordance test: results of a french pilot study

Authors:

Louis Sibert (louis.sibert@chu-rouen.fr)
Stefan J. Darmoni (stefan.darmoni@chu-rouen.fr)
Badisse Dahamna (badisse.dahamna@chu-rouen.fr)
Marie-France Hellot (Marie-France.Hellot@chu-rouen.fr)
Jacques Weber (jacques.weber@chu-rouen.fr)
Bernard Charlin (bernard.charlin@umontreal.ca)

Version: 2 Date: 12 July 2006

Author's response to reviews: see over
Dear Mrs Parkin,

Thank you for your e-mail dated June 26 and the useful comments made by the reviewers. In reply, please find enclosed our point-by-point reply to the Reviewers comments as well as a revised version of the manuscript.

Thank you for considering our article for publication in BMC Medical Education.

With best regards,

Louis Sibert, MD, Ph.D.
Reviewer N° 1: Carlos Brailovsky

1) The Reviewer’s comment: “The discussion is well written. However, I do not agree with one of the major conclusions. The authors suggest that the results showed a trend in the increase of the mean of groups with different clinical expertise (Table 3). When we look at the results presented in this Table, we don’t see any trend. Three out of four groups perform in exactly the same way. It is not necessary to analyze the results statistically to see that there is no difference among groups. The student’s performances show statistical differences compared to the urologists and chief-residents, but not when compared to the residents. Even if the authors try to explain these results by the size of the samples, they cannot support their suggestion that there are trends. Furthermore, the authors mention that the global correlation of scores with levels of training Observed supports the instrument construct validity. However they did not present any kind of results where this was demonstrated.”

- Authors reply:

From our point of view, this is a very important comment. We acknowledge that the body of evidence about construct validity of the test is not strong. Nevertheless, to our knowledge, our study is one of the first reported SC test applied to the Internet, our data are promising and have prompted us to optimise SC test content. This research is currently under development.

As suggested by the Reviewer, this specific point is now clarified. Sentences regarding arguments in favour of construct validity have now been deleted. It now reads in the Conclusions Section of the Abstract: “This Web site has permitted to quickly confirm reliability of the SC test and develop strategy to improve construct validity of the
test when applied in the field of urology.” In the Discussion section, it now reads: “As regards validity, our results showed that students obtained significant lower results than chief-residents and than urologists. Students performance was lower than residents performance but without any statistical difference. The lack of significance in the difference of scores between the other participants’ groups can be explained by the higher variability of residents and urologists’ scores. Despite the global correlation of scores with level of training observed in our study, the construct validity of the SC test online is not straightforward. Nevertheless, it is interesting to note that the less experienced group of participants obtained the lower results.”

2) The Reviewer’s comment: “I am not certain how the Cronbach alpha is calculated. To be able to interpret this coefficient, items must be mutually exclusive. But in the Script Concordance Test, even if they are independent of each other, are not independent of the domain of each clinical vignette. So, it is conceptually wrong to use the items as units of analyses but it is necessary to use the vignettes as units of analyses. The authors do not explain how they analyze their results, I suppose that they have used the items instead of the vignettes. The population under study is heterogeneous. The students’ results are lower than those of the other groups. This can increase the alpha coefficient artificially, making its interpretation hazardous.”

Authors reply:

This is a very interesting comment. In fact the definition of variables to measure reliability of a clinical competence assessment is not straightforward. No consensus exists in the literature regarding the use of items or clinical vignettes as units of measure.
for the calculation of Cronbach’s alpha reliability coefficient. Further research is warranted in this particular domain. In our study, we have used the items as units of analysis to assess reliability of the examination, in order to be representative of measures used in the majority of previous published studies regarding SC test. We agree with the reviewer that we should explain how we analyze the reliability coefficient. This is now mentioned in the Method-statistical analysis Section. We have added the following sentence at the end of the Method – statistical analysis Section: "In this study, the items were used as units of reliability analysis, in order to be representative of reliability measures used in previous published studies on SC test."

3) The Reviewer’s comment: “Finally, the example given in Table 2 is not very convincing in supporting the model of clinical reasoning. In fact, three items look as if they were from a MCQ and do not need the vignette to be answered. This was confirmed by two urologists that I consulted. I hope that the other vignettes and items used in the project behave differently. I would like to know if all the vignettes were similar. If not, I think that another example will be better

Authors reply:
We agree with this comment. As suggested by the Reviewer, the example of SC test section given in Table 2 has been modified. We now present a clinical vignette and items from the therapeutic section of the SC test in Table 2. This is now mentioned in the Method – Development of the SC test section. The sentence now reads: "An example of items from the therapeutic section is illustrated in Table 2.”
Reviewer N° 2: Clarence Kreiter

1) The Reviewer's comment: "On page 17 they say the SC test can be used in situations where there is no consensus among experts, it is in the literature or in daily practice?"

- Authors reply:

In fact, SC test can be used in daily practice situations. The commonly used format of clinical reasoning assessment tools (MCQ, written-simulations) requires to provide only one correct answer for each item by consensus among jury members. The scoring process of SC test offers the opportunity to assess clinical reasoning skills in contexts where evidence-based medicine cannot be applied. These are common situations in daily practice. The above mentioned sentence on page 17 has now been modified. It now reads: "The test can be used in situations where there is no consensus among experts in daily medical practice".

2) The Reviewer's comment: "I think it would be helpful for the authors to describe in what context a test composed of such items might be useful. They imply high-stakes exams on page 10. Do they suggest then using it in licensure or competency testing? It would be useful to discuss how these items might be used."

- Authors reply:

SC test can be useful for assessment of clinical reasoning in contexts of uncertainty. To date, this tool has been used in formative assessment contexts. Its psychometric properties make it possibly interesting for summative purpose. This is our next goal. The above mentioned sentence on page 10 ("...using a panel of 20 members should be recommended for high-stake examinations.") was in reference to research findings
which established that recruiting 10 or more members in a reference panel is associated with acceptable reliability, and using more than 20 members shows only a marginal benefit in terms of psychometric properties. In order to make it clear, we have modified the following sentence on page 10, in the Method Section, paragraph Participants. It now reads: "...using more than 20 members shows only a marginal benefit in terms of psychometrics properties."

3) The reviewer’s comment: "Page 5: The authors state most PBAs are only a measure of behaviour. Do they implying that measuring behavior is less useful than measuring a precursor to behavior? Perhaps they meant to imply that current PBAs are measuring trivial behavior. If so, they might be right, but they need to elaborate."

- Authors reply:
Professional competence is a multidimensional entity, which comprises of behaviour and cognitive skills and cannot be adequately measured by a single assessment method. Performance-based assessment, such as Objective Structured Clinical Examination, is considered the Gold Standard for assessment of observable skills. This is what we refer in the text as "behaviour". SC test only explores the capacity of data collection in the making of clinical decision and should be used in complement with other assessment tools of clinical competence. As suggested by the Reviewer, this point has now been clarified. Page 5, background Section, it now reads: "Furthermore, most current performance-based methods of professional competence assessment (e.g. Objective Structured Clinical Exams) are only measures of observable clinical skills."
4) The Reviewer’s comment: “Page 7: I am unsure why the diffusion of the SCT on a large scale will permit the researchers to confirm the utility as a strategy for investigating the process of decision making! However, it is reassuring that here they seem to imply that the SCT is a tool for researching or investigating decision making rather than educational assessment.”

- Authors reply:
We agree with this comment. In fact, the major interest of SC test diffusion on a large scale was to permit us to assess psychometric qualities of the test in a larger population than previous published studies, moreover in a relatively short period of time. This is now clearly mentioned in the text.

5) The Reviewer’s comment: “Page 10: Why do the authors seek a norm referenced database for the reference panel to derive a scoring key? For example seeking a broad range of training levels makes little sense. I am not sure, but if I understand what they seek to do with an SCT aggregate scoring key, they should rather characterize a group of experts with consistently high levels of experience and training.”

- Authors reply:
This is a very interesting point concerning the methodology of SC test scoring. We attempted to find a norm referenced data base for the reference panel. For a SC what we referred to as “experts” were experienced urologists. Members of reference panel should reflect all types of clinical practice in urology. This point is now clearly mentioned in the text.
6) The Reviewer's comment: "The new territory described in this research report relates to the new SCT web-based testing software. However this article and research design does not focus on evaluating or describing the impact of the new delivery mode."

- Authors reply:
We totally agree with this comment. In fact, as mentioned in the article, the goal of this study was to quickly include a large population of candidates via the Internet. The impact of the Internet as a delivery mode for clinical competence assessment remains under study. Further research is warranted on this particular topic.

7) The Reviewer's comment: "Other concerns relate to the fact that 45% did not complete the exam. How were missing values scored? Was the percentage constant across training levels? If not, could the differences in scores by level be explained by level of volunteer examinee commitment to performing well on the exam? This is an especially important question since it was only students who scored significantly different and other groups."

- Authors reply:
As regards missing values scored, we stated on page 13 that: "statistical analysis was performed only on the participants who fully completed the test." In cases where the participant did not complete the SC test, he was not given a score. The proportion of participants who fully complete the test was constant across training levels; therefore, we do not think that differences in scores by level can be explained by the percentage of candidates who did not complete the test among groups of participants with different levels of experience in urology. Nevertheless, as suggested by the Reviewer, we have added (page 13) the different percentage of candidates who fully completed the test. We
The authors are grateful for the useful comments made by both reviewers.

Thank you for considering our revised manuscript for publication in BMC Medical Education.

With best regards,

Louis Sibert, MD, Ph.D.