Author's response to reviews

Title: The Use of Standardized Patients for Mock Oral Board Exams in Neurology: A Pilot Study

Authors:

Brett Kissela (brett.kissela@uc.edu)
Steven Harriss (harrissd@upstate.edu)
Dawn Kleindorfer (dawn.kleindorfer@uc.edu)
Christopher Lindsell (christopher.lindsell@uc.edu)
Robert Pascuzzi (rpascuzz@upui.edu)
Daniel Woo (daniel.woo@uc.edu)
Jerzy Szafarski (jerzy.szafarski@uc.edu)
Daniel Kanter (daniel.kanter@uc.edu)
Alexander Schneider (alex.schneider@msj.org)
Michael Sostok (michael.sostok@uc.edu)
Joseph Broderick (joseph.broderick@uc.edu)

Version: 2 Date: 9 February 2006

Author's response to reviews: see over
Response to Reviewers:

Title: The Use of Standardized Patients for Mock Oral Board Exams in Neurology: A Pilot Study
Version: 1

We appreciate the thoughtful reviews of our manuscript. Hereafter are our responses to the critiques from each reviewer.

Reviewer #1:

General:
The authors describe a 2 site pilot project of the development and analysis of a new standardized patient live patient evaluation for neurology trainees. They conclude that this is highly promising and further study is required. The pilot described demonstrates a highly innovative and very useful starting point for institutions and agencies to consider how neurology residents can be evaluated to assess competency in numerous areas.

Thank you.

Minor Essential Revisions:
1. Page 4 states that the manuscript describes the experience at UC and the abstract says UC and IU. IU should be added to this statement as this is confusing to the reader.

The exercises were conceived and performed at UC, but IU residents were invited to participate in the first year’s experience given the formative nature of this exercise (see response to Reviewer 2, item #8 below). The reference to UC has been deleted.

2. Clarification on whether this was voluntary for trainees at both institutions is required.

This exercise was required for trainees at the University of Cincinnati, taking the place of the yearly “mock oral boards”. The exercise was voluntary for trainees from Indiana University. A statement to this effect has been added to the text.

3. Clarification if this was considered human subjects research at both institutions is required. If so, details of consent are required.

This exercise was developed to be part of the UC Neurology Residency educational program, and was not developed for research purposes. Resident evaluation forms were developed for internal use in evaluating resident performance over time. Surveys were developed to determine if the experience was beneficial for residents and if the SP simulation was comparable to the use of “real patients”.

Thus, this was not considered human subjects research. This was discussed with the IRB at the University of Cincinnati prior to the exercise described, and they agreed that IRB review was not necessary. Given that the reported exercise occurred in 2002 and the environment of human subjects research has evolved over time, this was again
discussed with the IRB in preparation for this response. Once again, they agreed that IRB review or exemption was not necessary.

Per 45 CFR 46.101(b), the following are not considered human subjects research (emphasis added by authors):

"1. Research conducted in established or commonly accepted educational settings, involving NORMAL EDUCATIONAL PRACTICES, such as (i) research on regular and special educational instructional strategies, or (ii) research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.

2. Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), SURVEY procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects’ responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects’ financial standing, employability, or reputation."

It should be noted that the resident physicians who participated cannot be identified from the information presented.

Several clarifying statements have been added to the text.

4. **Weaknesses of this pilot should be discussed. The lack of 2 evaluators on every subject, the feasibility issue as it was suggested that faculty had trouble getting to exam, the 3.5 score re inaccurate portrayal of findings (between neutral and disagree) should be discussed.**

We have included the lack of 2 evaluators as a weakness of the pilot in the discussion section. Obtaining 2 faculty evaluators was indeed a logistic challenge that may make this exercise impractical for some programs. However, the inability to obtain 2 faculty evaluators was partly due to inexperience of the lead author with organizing such exercises, and this obstacle has been overcome in subsequent examinations.

With regard to the latter issue (3.5 score re inaccurate portrayal of findings), the ability of the SP’s to portray a neurologic patient was one of the major questions to be answered and so this was assessed with multiple questions on the survey. Both residents and faculty were asked if the history and physical exam were realistic (the first two questions in Table 3. Participating faculty members were also asked whether the residents were distracted by inaccurate portrayal of the physical findings. The 3.5 score on this latter question, indicating that the residents were not distracted by inaccurate portrayal of physical findings, is consistent with the earlier faculty response that the exam portrayal was realistic. We would be willing to clarify further in the discussion section if the reviewer feels that this is necessary.

5. **The stats section requires more detail. It is unclear what the authors were comparing using t-test as numerous scoring instruments are discussed. Clarification of which scores were compared is required. Brief mention in analysis**
of the summary stats for the other scoring instruments is required. Pearson’s correlation is described in the results and not methods.

The statistics section has been revised according to the reviewer’s comments. The student’s t-test was used to test for differences between faculty and resident responses on each survey item. The statistical analysis of the summary stats for resident performance has been clarified. A sentence with regard to Pearson’s correlation coefficient was added to this section.

Please also see response to Reviewer 2, item # 9.

Discretionary Revisions (which the author can choose to ignore)
1. References 2-19 are excessive - several that are representative would be recommended.
   The reference list has been revised. We have deleted many of the previous references (#’s 4, 6, 9-11, 13, 15, and 28 from the original list). We repeated a literature search and review, and have added one new reference showing reliability and face validity of “geriatric” SP exams for medical students, residents, and fellows (Nagoshi, et al). Please also see the response to Reviewer #2, items #1, 4-5, as 2 additional references were added from Hodges, et al.

2. Clarification on feedback after the "dress rehearsal" would be helpful. Did any require retraining and reevaluation?
   Minor points were clarified with the SP’s with regard to history and portrayal of physical exam findings. There was no need for substantial retraining or re-evaluation. A statement to this effect has been added to the text.

Reviewer #2:

General
1.) This manuscript documents the development and implementation of a standardized patient-based practice oral exam for neurology residents. The authors provide significant details about the cases and checklists developed for the practice exam. Unfortunately, the results as reported are limited because of the small sample size of participants in the pilot test. The meaningfulness of the results presented as means and SDs is questionable because of the small sample size.

We agree. We have clarified in the discussion that our intention was to present our innovative methodology (see response below to Reviewer 2’s item #6) and not to draw meaningful conclusions about the reliability or validity of this exercise. However, as with any pilot study, it is important to examine the data collected to see if any major trends are detected. We have attempted to do so in this manuscript, but did clarify that we agree with the reviewer. We have now explicitly stated in the discussion that the statistical results are not highly meaningful (and should be interpreted with caution) due to the small sample size.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)
1.) While the literature review provides many citations to published work, there is no synthesis of the published literature to provide a context of why the current study is important or specific issues related to oral examinations that would need to be considered in this paper.

The exercise was created in response to the real-world situation of testing clinical skills within neurology residency training as described in the introduction. We developed this exercise to apply the use of standardized patients to the common neurology residency practice of yearly mock oral boards. This is a novel use of SP’s and a literature search found no relevant references with regard to the use of SP’s in neurology training for residents. This question has relevance to the evolving manner in which neurologists seek board certification from the ABPN.

Our literature review (references 2-28 in the original reference list) was intended to show that the use of SP’s in medical education is widespread and successful, at both the medical student and resident level. Thus, we hoped to demonstrate that our project is relevant and timely. We have revised the reference list—please see response to Reviewer #1, item #1 (Discretionary revisions) above. Please also see the response to items #2, 4, and 5 below.

2.) There is no clear research question or hypothesis provided to frame this investigation.

We have clarified the questions to be answered in the final paragraph of the introduction as follows: “Our broad objective was to determine if this exercise was a practical and useful alternative to utilizing actual patients for mock oral boards. We specifically wanted to determine if SP’s could successfully portray neurologic patients to the satisfaction of the residents and faculty involved in the exercise.”

3.) The authors fail to address many issues related to oral exams such as case specificity and examiner variability. They do suggest on page 10 that the fidelity of the SP over time is a concern, however they provide no evidence from the literature review to support this as a concern.

Case specificity in the SP literature usually refers to the variability in the “person x station” interaction. Relevant to our manuscript, we have only one case per level of training. Each resident only examines one SP; this is not a multi-station exercise. For each SP case, the same examiners were utilized for all exams at that level. This reduces examiner variability, but further limits the ability to interpret the summary results, such as in the survey where the opinions of the faculty members on their specific stations are combined and generalized to the entire exam. We agree with the reviewer, and would be willing to add a comment to the discussion (as a limitation) if they feel strongly, and yet we feel that this might confuse the reader. This limitation is inherent in the statement about how small sample size must temper the reader’s ability to draw meaningful conclusions about the data presented.

With regard to examiner variability, we have explicitly added this to further clarify that the lack of 2 examiners for each exam is a major limitation of the study. Please also see the response to Reviewer 1, item #4.

4.) The work of Hodgins and colleagues might prove useful in framing possible research questions: An objective structured clinical examination for evaluating psychiatry clinical clerks (Academic Medicine, 1997) and Validation of an objective structured clinical exam in psychiatry, (Academic Medicine in 1998).
These references have been included in the revised references as representative of evaluation of medical students (1997 and 1998 papers) and residents’ clinical skills (1998 paper) with SP’s. The 1998 paper is especially valuable in its examination of validity with such exams.

5.) Another source that might be helpful is Kumar et al, Locating, characterizing and minimizing sources of error for a paper case-based structured oral examination in a multi-campus clerkship. (Advances in Health Science Education, 2001).

This is an excellent article with regard to methodologic issues in validating clinical skills assessments in clerkships. However, we did not feel that it was relevant to this paper given the focus on SP utilization, and was thus not included to keep the reference list shorter and representative (please see response to Reviewer #1, Discretionary Revisions, item #1).

6.) The authors suggest that the use of standardized patients in neurology oral practice exams is a novel application, and this is indeed the strength of their paper. This paper would be stronger if it was rewritten in the form of an educational innovation rather than a research report. The focus would be on the issues that this innovation in assessment would address. The appendices provide valuable detail about the development and implementation of this innovation. In addition, the faculty and resident questionnaires can be reported as preliminary evaluation data about the innovation. The literature review would need to provide a context for the problems this innovation is designed to address.

We agree with the reviewer. It was our intention from the outset to present this exercise as an innovation for neurology residency training education and not as a substantial research report; hence the term “Pilot Study” in the title. However, as stated above in response to the reviewer’s general comments, we feel that it is important to examine the data collected in this pilot study—these are preliminary data. Had the data indicated that this exercise was a failure, further efforts would not have continued and we would have concluded that the use of SP’s in simulating neurologic disease was not worthwhile. Thus, we feel it important to keep the data in the manuscript. We have modified the language of the discussion to emphasize that these are preliminary data from a pilot study. In summary, we do agree with the reviewer and hope that edits to the introduction and discussion will frame the question appropriately to match our original intent. Indeed, one of our objectives in submitting this manuscript was to have the detailed cases and forms published in the literature so that others might utilize them and provide further data with regard to the reliability and validity of the exercise.

For discussion of how the literature related to the context of this exercise, please refer to our response to Reviewer 2, item #1 above. The use of SP’s to portray a neurologic patient (performed at a level that would appear accurate to neurology residents and fully-trained neurologists) is a novel application. It is our hope that future work will establish this exercise as reliable and valid, adding to body of evidence supporting SP use in medical education.

Other concerns related to this manuscript as it was submitted for review: 7.) There is no evidence of IRB approval for this study, or that it was submitted for IRB approval.
Please see above response to Reviewer #1, item #3.

8.) The authors should clarify in the text if this practice oral examination was implemented as a formative or summative assessment for the residents.

For the residents, this was implemented primarily as a formative evaluation. The yearly mock oral boards provide experience with the current format used by the ABPN for neurology certification, and feedback from faculty evaluators (verbal and written) provides useful information to improve clinical performance and future test performance. The previous format involving real patients was a formative evaluation, as is the SP exercise described in this manuscript. This has been added to the introduction.

9.) In the Results section, the authors report the use of inter-observer agreement; the use of Cohen’s Kappa coefficient should also be presented. (Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960; 20(1): 37-46. see also Landis, J.R. & Koch, G.G. The measurement of observer agreement for categorical data. Biometrics. 1977; 33: 159-174.)

We agree that better measures of concordance are available than mean differences and correlations. We have added Lin’s concordance correlation coefficient as the most appropriate concordance measure for the score, which is a continuous variable. We originally elected not to use a concordance measure due to the small sample size (n=6); Lin’s concordance correlation coefficient has not been shown to be robust when sample size is fewer than 10. We assumed the reader to be able to determine strength of concordance based upon the raw differences for this small sample.

It should be noted that we considered using Cohen’s kappa to assess concordance for items making up the scale. There are several problems with this. First, no reviewer pair was the same for any resident evaluation. Second, two reviewers elected to use ‘half’ point measures on the scale (i.e. we provided an integer scale from 1 to 5 and they responded with 2.5 or similar). This places doubt on the value of chance corrected agreement since not all reviewers used the half-point measure. Finally, based upon the small sample size we were reluctant to compute a kappa for each item, and computing kappa assuming independence of items was not necessarily appropriate. For the reviewer, the overall Cohen’s kappa was 0.841, which was calculated ignoring these limitations.

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)
None

Discretionary Revisions (which the author can choose to ignore)
None