Author’s response to reviews

Title: Validation of the Imperial College Surgical Assessment Device for spinal anesthesia.

Authors:
Marcia Corvetto (marciacorvetto@gmail.com)
Andrea Araneda (aranedaandrea@hotmail.com)
Carlos Fuentes (cfuentes812@gmail.com)
Pablo Achurra (achurrapablo@gmail.com)
Pablo Miranda (pablofmh@gmail.com)
Fernando Altermatt (fernando.altermatt@gmail.com)

Version: 1 Date: 19 Jun 2017

Author’s response to reviews:

June 19th 2017

Dear Dr. Schaefer,

Thank you so much for the opportunity to revise the manuscript. Your helpful comments and suggestions have been essential to improve it. We have addressed each of your concerns systematically, reiterating each comment and following it with our response.

Manuscript revisions have been highlighted in color, as requested.

Point-by-point response to specific comments:

Lidia Mora Miquel, M.D. (Reviewer 1):

1. - The novel point would be to integrate the simulation in the learning process of anesthetic abilities, with the intention of discriminating if by using this tool of validation of skills could be detected the sufficient degree of competences acquired so that the residents of anesthesia could carry out some actions without being supervised. I believe that this intention should be clearly stated because it is glimpsed in the text during the discussion.

Thank you for the input. We have included this idea in the discussion section.
2. - The purpose of the study describes the intention of the discrimination of levels of proficiency between operators in order to validate a tool for the training with simulation techniques. The practical aim could be added in my opinion, since the goal in medical education is to improve the entrustable professional activities (EPA). The authors declare this is the first study using this educational methodology performing a spinal anesthesia, so it strengthens the idea of an advance in medical education.

    We have added this idea in the discussion section.

3. - The abstract is too long. Too many data and of course, not to mention the tables inside the results. It should be abbreviate. Typographic error when writing underestimate.

    We have edited the abstract. We have shortened it to 277 words. It is not clear for us what did the referee try to say with the sentence “Typographic error when writing underestimate”. We’ve proofread the manuscript, correcting the typos we found.

4. - It should be described whether the distribution of cases among the three groups was randomized or not because it could not lead to a significant selection bias given the difference in experience between participants.

    Participants were not randomized, because the idea was to have 3 different levels of expertise. Construct validity intended to determine the ability of the assessment tool to discriminate between operators with different experience level. In our study, each level of experience was established a priori, and participants were recruited as part of their specific group.

5. - The description of the validated Global Rating Scale and Checklist (Appendix 1 and 2) are missing. Please, provide the documentation or describe them in the text.

    We have attached both documents as Appendix 1 and 2.

6. - The significance of the moderate negative linear correlation result of the Spearman correlation analysis between total path length measured with ICSAD could be explain at the end of the results.

    We have added the explanation at the end of the results.
7. Inside the discussion another explanation of the aim of the study is described, just before the limitations. In my opinion it could be added to the introduction of the manuscript and then discussed again at the end if it considered appropriate.

We have added this idea in the introduction. We have maintained the paragraph in the discussion section, as well.

8. The main limitation of the study is that it is probably underpowered because of the reduced number of cases. The authors justify that problem referring to previous publications but in order to establish conclusions it should be a larger sample of cases.

In fact, previous literature has used smaller sample sizes. For example, Van Sickle used 5 novices and 5 experts for a validation process. [1]

Our explanation of the lack of significant difference between experts and intermediates is because experts and experienced non-experts (third year residents) have similar performances during an average-difficulty spinal anesthesia.

In order to support this explanation, a post hoc power analysis for skewed distributions was done.

9. The justification of the validation of the ICSAD is well conducted and the reasons to choose a simulation model instead of real patients is correct and well explained. I think it could be remarked that this is a first experience with the simulation model and ICSAD performing spinal anesthesia and the authors could insist in the importance of the new educational methodology based on simulation training in Anesthesia and other disciplines.

Again, thank you for this input. We have added this idea in the manuscript.

10. The references in the bibliography could be normalized and standardized following the rules of the journal.

We have edited the bibliography accordingly.

11. Table 1: add (y) to the age variable.
We edited the table.

Dan Benhamou (Reviewer 2):

1. - The authors state that the methods used allow for both construct and concurrent validity. The authors should however add one or two sentences in the Methods section to explain why is meant by these two terms which are likely unfamiliar to most readers.

   Thanks for the input. We have added the explanation of construct and concurrent validity in the methods section.

2. - One limitation is related to the absence of power analysis done before the study. It is noticeable that the number of subjects included is small at first glance and several results may remain insignificant simply because of inadequate power. Results however show that this small number is enough to obtain several interesting correlations.

   We did not do a power analysis a priori. We tried to find previously published data such as total path length (TPL) of experts and novices in order to guide this calculation. Essentially, there are just two studies to get information to do a sample size calculation. Hayter presented a validation process in epidurals. They concluded that novices took more time, longer distances and more movements, but they did not present a table detailing those dexterity scores. Hayter’s Figure 1 shows the dexterities of each participant, but it is difficult to extract global values. Figure 1 shows TPL values, ranging from 300 to 1000 mm in the expert group.[2] Therefore, we do not have median or mean values to work with. On the other hand, Chin presented a validation process in supraclavicular blocks. The needling hand shows a distance travelled of 154 meters in novices versus 92 meters in experts. With those so dissimilar data, it was really difficult to have accurate information to perform a proper sample size calculation or power analysis.[3]

   In this context, previous literature had used smaller sample sizes. For example, Van Sickle used 5 novices and 5 experts for a validation process. [1] With all these considerations in mind, we decided to use 30 participants (10 per group). Our main explanation of the lack of significant difference between experts and intermediates is because experts and experienced non-experts (third year residents) performed similar during a regular (standard, non-challenging anatomically difficult) spinal anesthesia. We have tried to expand this explanation in the discussion section. In order to support this explanation, a post hoc power analysis for skewed distributions was done.
3. - The post hoc analysis is mentioned in the Discussion section but not described in the Results. Please provide details in the Results section.

   Accordingly, we have added the explanation in Methods and in the Results Sections.

4. - The authors who described the system do not call it ICSAD but rather synchronised video and motion analysis. Please modify also in the title.

   Thanks for the observation. We have eliminated from the title. We have maintained the acronym in the manuscript in order to save words.

5. - P9, L120: is the Gaumard S411 a model for "easy lumbar puncture"?

   The name of the Gaumard S411 is “Lumbar puncture trainer”.

   [http://www.gaumard.com/s411](http://www.gaumard.com/s411). It is a simulator designed for resembling a standard level of difficulty procedure. It does not consider any anatomical variations increasing the difficulty of the procedure (obesity, hyperlordosis, scoliosis, etc).

6. - P9, L 127: Because the device is not well known, it would be useful to add one or two pictures showing the system and how are placed motion capture devices on the hands.

   Thanks for the comment. We have added a picture showing how the sensors are placed on the hands.

7. - Ref 12 should be replaced by a reference which can be more easily accessed by readers such as Dosis A et al 2005 and/or Aggarwal R et al, 2008.

   We have replaced the reference.

8. - P12, L168: withdraw the word "perfect" and only provide the kappa coefficient here. The word "perfect" is a comment and this has no place in the Results section.

   Agree. We have edited that part.
9. - P16, L245: the sentence starting by "Also, Hayter..." is not clear. Please modify.

   We have edited that part.

10. - P16, L258: replace "by" by "for".

   We have edited that part.

Klaus Hofmann-Kiefer (Reviewer 3):

1. - Novelty: Unfortunately, a very similar study concerning the applicability of ICSAD to discriminate technical skills of various operators to perform an epidural anesthesia in pregnant women has been published in 2009 by Hayter MA et al. (1). Both studies included the same number of subjects, used the same methods and statistics and reported nearly identical results (see below). They also present some similarities in the discussion section.

   We are totally conscious that Hayter published previously a study with a similar methodology in epidural anesthesia in pregnant women. However, there is an ongoing trend to establish more objective measures of technical skill, the referee has to conceal that describing the patterns for a specific procedure (Hayter’s epidural or Chin’s supraclavicular) does not automatically transfer the data and the interpretations to any other type of procedures performed by a given specialty. From our point of view, each procedure should have their own characterization. For example, number of movements and distance travelled by hands should be established in order to have validated thresholds to guide training. In addition, spinal anesthesia is a different procedure than PNB and epidurals because they present different steps to perform the procedure and different learning curves, as it was described by Konrad et al.[4]

   • Regarding number of subjects: Previous studies in the surgical field have used 5 to 8 participants per group.[1, 5] Hayter used 8, 9 and 12 participants per group. Van Sickle used 5 novices and 5 experts for a validation process.[1] Following this reference we decided to enroll 10 subjects per group.

   • Methods: 3 groups were used for the validation process. Construct validity intended to determine the ability of the assessment tool to discriminate between operators with different experience level. Although Chin used 2 groups, we thought that 3 groups would do a better job illustrating the construct validity.

   • Statistics: Hayter used MANOVA. We used Kruskal-Wallis test.
• Results: Hayter affirmed that novices took more time, longer distances and more movements, but they do not presented a table with those dexterity scores. They did not present median or average values.

• Discussion has changed with the contribution of reviewers.

In conclusion, although Hayter’s work and ours share several aspects in terms of aims and methodological design, they are different in terms of the procedure studied and the results observed, probably because the levels of difficulty observed in epidural and spinal blocks are different, and therefore, would be worthy to describe them independently.

2. - Methods provided in the current study are principally adequate and sufficient. However, there are far reaching similarities to the study conducted by Hayter et al.: In both cases study groups consisted of a novice group, an experienced resident group and an expert group of attending anesthesiologists.

In both studies technical skills were evaluated by the ICSAD system and through videotapes which were judged by two independent, blinded observers.

In both of the studies the reliability of the ICSAD was compared to a GRS and a task specific checklist.

As we previously stated, there are many similarities between our study and the one done by Hayter et al. Both are concurrent and construct validity studies using ICSAD, but for two different procedures. It is difficult to argument against any absence of novelty in methodological aspects, since that many studies in different subjects can share similar methods. Let us see how do these similarities go:

• Study groups: They were the same groups (novice residents, experienced residents and expert anesthesiologists). From our standpoint, the selection of these groups is the better way to test the tool in discriminating expertise. Other studies have done their validation process with only 2 groups. However when we designed this study, we thought that 3 groups would be better in terms of illustrating the construct validity of the instrument.

• Assessment tools: ICSAD and GRS were the same in both studies. Checklist was different, because the steps required for each procedure are different. We used a specific checklist designed for subarachnoid blocks.

• Videos: Both studies used videos to apply the observational tools. From our point of view they are better than real life assessment, because it allows to retrieve more information from the videos, pausing them and reviewing them as many times as it is needed.
• Concurrent validity: In both of the studies the reliability of the ICSAD was compared to a GRS and a task specific checklist.

There are several points where we have tried to be more informative than previous studies:

• Results: Hayter et al’s results show that novices took more time, longer distances and more movements than experts, but they do not show a table with those dexterity scores. There are no median or mean values from the groups to compare with. Figure 1 shows the dexterities of each participant, but it is difficult to extract cut-off values from it. Also, Figure 1 shows a total path length (TPL) ranging from 300 to 1000 mm in the expert group.[2]. These values are somewhat intriguing. From our point of view, distance travelled during a peridural puncture should be more than 1000 mm (1 meter). Surprisingly, Chin reported 92 meters of the needle hand in experts, performing a single shot supraclavicular block.[3] Obviously, considered all together, these results are inconsistent in terms of the expected TPL for these two procedures. In this context, we intended to better characterize the data of a spinal block and we have presented the data more comprehensively than previous studies. We have tried to present specific cutt-of values, in order to have published data for each procedural skill in anesthesia. We divided the procedure in preparation and needling phases. We presented median and range values, in order to have evidence of the exact distance that hands travelled through this procedure.

• Statistics: Hayter wrote “all of our primary endpoints, the ICSAD, the checklist, and the GRS were analyzed parametrically using a multivariate analysis of variance (MANOVA)”. Significant differences were analyzed using a Tukey’s post hoc test. In our study a nonparametric 1-way analysis of variance (Kruskal-Wallis test) was made; subsequently, Dunn's correction test for multiple comparisons was performed in order to determine the different pair.

3. Unfortunately the appendix containing the detailed GRS and the checklist of the current study were not included in the PDF sent to the reviewer. However, at least the GRS seems to be the same in both of the studies.

   Thanks for mention it. We have attached both documents to the file.

4. In contrast to Hayter, Corvetto used a simulated torso to test the skills of their subjects. In the eyes of the reviewer this facilitates comparability and can be seen as an advantage of the current investigation.
Thanks for the comment. We decided to perform our study using a simulator for two reasons: First, due to ethical and safety concerns. Traditional approach for training residents implies that, if necessary, a permanent supervisor provide feedback or even intervene during the procedure affecting crude measurements and its subsequently validity and interpretation. Second, using simulators standardize the level of difficulty of the procedure for every operator.

5. Both studies use the same statistic methods which are - as far as I can say - appropriate. Unfortunately Corvetto et al. do not provide a statistical analysis concerning demographic differences between their study groups (Table 1). On first sight, it seems possible that "intermediates" and "experts" do not differ significantly with regard to the number of spinals performed in the last month/last 6 month. This may have contributed to the fact that the ICSAD was not able to differentiate between these groups.

Thanks for the comment. We have done a statistical analysis concerning demographic differences between the study groups. A Chi-square test was used to compare categorical variables and continuous non-normal variables were compared using a Kruskal-Wallis test.

6. Both studies have a major statistical flaw concerning power. However, unlike Hayter et al., who simply stated that their study might have been underpowered concerning the most important outcome parameters (ability of the ICSAD to differentiate between groups), Corvetto et al. provide a "post hoc" power analysis. Despite the well-known fact that post hoc power analyses are questionable in general (2), the calculations behind this power analysis are unclear (lines 250-254) and primary outcome parameters are not defined. According to Corvetto, an effect size of 1.5 SDs (of which parameter??) led to a power of 88% (?). However, SD is questionable for calculating power when comparing not-normally distributed data, what is obviously the case in the current study (see line 155).

Actually we did not do a power analysis a priori, for the reasons previously described: We tried to find previous published data such as TPL of experts and novices in order to guide this calculation. Hayter presented a validation process in epidurals. They affirmed that novices took more time, longer distances and more movements, but they did not present a table with those dexterity scores. Therefore, we do not have median values, ranges or any other consistent data to perform a sample size calculation. As a matter of fact, Hayter’s Figure 1 shows the dexterities of each participant, but it is impossible to extract cut-off values from it. The same figure shows a total path length (TPL) ranging from 300 to 1000 mm (0.3 – 1.0 meters) in the expert group.[2] Intriguingly, Chin et al reported way too different values for TPL in a supraclavicular block. The needling hand shown a distance travelled of 154 meters in novices versus 92 meters in experts. With such dissimilar data, it is impossible to perform a sample size calculation or any a priori power analysis.[3]
In order to facilitate a Post Hoc power calculation for skewed distributions and to determine the effect size of the study, a rank estimation of TPL (rank-TPL) was performed for each category. We choose TPL because it is the dexterity variable that is best correlated with GRS scores. Then, we tested this transformation with Shapiro-Wilk being able to assume normality. In this context, the observed effect size is 1.5 SD for rank-TPL and the study findings are retained when analyzed with One-Way ANOVA. Therefore, with a sample size of 10 subjects for each category and a significance level of 0.05 (two tailed) we calculated a power of 86% to detect a difference.

7. At least the authors used non-parametric statistical tests (3). The authors claim "that there are no previous data available to calculate a sample size on "a priori" basis for this kind of studies but others have been done with 20-30 participants". However, in the eyes of the reviewer it shouldn't have been too difficult to define a proper outcome parameter characteristically for the ICSAD (for example path length), then to define a proper effect size and to perform a valid power analysis.

We agreed that we could define a proper outcome parameter to calculate power. Unfortunately we already did not do it when we designed the study. We have improved our post hoc analysis in order to answer the question regarding the possibility to be underpowered.

8. Results: Corvetto (as well as Hayter) found that the ICSAD was able to differentiate between novices and experts, but not between intermediates and experts. However, both authors do not blame this fact on the method itself but speculate about the possibility of too far reaching conformities between the latter groups. In the eyes of the reviewer it would have been more useful to simply stay with the facts: Obviously the ICSAD is not suitable to find differences between trained and expertly trained anaesthesiologists, but only between novices and experts. This result questions the method in general and should have been properly discussed. If reasonably worked up by Corvetto, this could have been a possibility to demarcate the current study from the very similar findings of Hayter et al.

We did not detect significant differences between intermediates and experts regarding total path length, number of movements and time. As we stated in the limitations section, a possible explanation is that most of the third year residents have already flattened their learning curves achieving enough proficiency to perform an “average” spinal anesthesia unsupervised. So finally, the problem is not the inability of the tool to discriminate, rather "intermediates" and "experts" do not perform significantly different in a simulated spinal anesthesia. In other words, third year residents perform not as “intermediate” operators but as “journeymen” when they are evaluated against a standardized spinal block simulator. He simulator does not discriminate between the performance of third year residents (high level of proficiency under standard
conditions) and experts. Maybe a simulator with capacities to modify the level of difficulty, incorporating anatomical variations such as obesity or axial deviations, could differentiate better between these two groups.

A previous study done by our group in central venous catheter insertion, showed significant differences between experts and intermediates, maybe because it is a more complex procedure (Corvetto et al. "Simulation-based training program with deliberate practice for ultrasound guided jugular central venous catheter placement." Acta Anesthesiologica Scandinavica, In Press, 2017)

From our standpoint, this study represents a contribution, because the use of this motion device in the evaluation of motor skills allows obtaining quantitative data complementing previously validated visual scales. Hayter did not report any median or mean values of an epidural installation. The simulated model (our study) gives us the opportunity to have real values of novices, doing the procedure freely, without the constraints imposed by ethical considerations in terms of the risk of improperly performed procedures in patients.

Having as many instruments as possible for evaluating motor skills could improve the learning process. In the future, if we want to setup metrics or cutoff scores to be achieved with motor skills training, a previous standardization of parameters to be used should be established for each procedure.

9. - Corvetto also compared the ICSAD results (only path length) to a validated GRS and obtained a correlation coefficient of -0.467. In contrast to Hayter's study the other modalities (duration, number of movements) of the ICSAD were not correlated to the GRS. This should be made up in a revised version and the results should be discussed, too.

Agree with the comment. We presented a table with all the correlations. We have improved the explanation of the correlation coefficient, as well.

10. - The rather moderate correlation between the ICSAD (path length, and if provided the other modalities), which does not find further mentioning, should also be a topic of discussion. GRS-testing and checklist-testing are useless, if the results of these tests are neither correlated to the ICSAD nor properly worked up.

Thanks for the comment. We improved the discussion of the correlation between tools.

11. - Line 51: this study aimed to …. (Change to simple past)
We have edited the abstract.

12. - Line 90 and line 98 (and others): The positioning of the literature cross references is inconstant: Please set all of them at the end of a sentence (after the dot) or in between, according to the author's instructions of the journal.

We have edited the references. We have set all them at the end of a sentence.

13. - Line 102: …study was to determine …..

We have edited this part.

14. - Line 104: this aim was only partially fulfilled (see General considerations)

We have edited this part.

15. - Line 137: Did the participants wear gloves? I think in some cases it should be easy to differentiate between the hands of a young unexperienced novice and an elder expert.

We have edited this part.

16. - Line 136 - 148: The reviewer does not understand why the authors undertook these complex preparations concerning GRS, if the results were hardly put into relationship with the ICSAD results.

GRS has been widely validated to assess different procedures, but it requires training for optimal reliability.[6] We tried to do our best in the use of this assessment tool.

17. - Line 153: Why were the results displayed as median and interquartile range? Normally this is the case if data are not-normally distributed. Which tests were used to verify data distribution?

Given the number of cases of this study we assumed a priori a not normally distribution. In this context, we decided to present the data as median and interquartile range. Supporting this
idea a Shapiro-Wilk test was done for default in SPSS when we performed the descriptive stats, given a not normally distribution.

18. - The post hoc power calculation and its methods/variables should also be displayed in the methods section.

We have improved the Post hoc calculation. We have mentioned it in methods and explain it more deeply in results (accordingly with the request of another reviewer).

19. - Line 170: According to the guidelines of Landis and Koch a kappa coefficient of 0.76 is "substantial", not "almost perfect" (only values > 0.8)

Thanks for the comment. We have edited this part.

20. - In line 168 to 170 the authors describe the inter-reliability of the GRS scores, but than do not describe the results of the measurements but suddenly change to the results of the ICSAD?? (Line 171). That's disturbing. Where are the GRS results of the preparation phase displayed?

We have edited this part. We have deleted the statement "almost perfect".

We divided the procedure in 2 parts for ICSAD dexterities. GRS and checklist are for the entire procedure (it has one point to assess preparation). We have edited the text in order to be more precise on this subject. We have separated the data of visual scales into another table.

21. - Line 181 and Figure 1: Figure 1 (as well as figure 2) does not provide any statistical data. The reader is not able to recognize whether there are differences between groups or not. The figure legend is insufficient.

Statistical differences are expressed in the tables. The idea of the figure is to have an idea of each participant value, the medians and ranges. We have improved the figure legend. We are open to include the statistical data into the figure, too if the referee considers it necessary.

22. - Line 190-193: Why was only path length correlated to the GRS??

We have added a table with all the correlations.
23. - Line 195: This conclusion is not valid, as you do not succeed in differentiating between intermediates and experts (ICSAD) and you only noticed a moderate correlation to the (validated) gold standard GRS.

Regarding the fact that we do not succeed in differentiating between intermediates and experts, construct validity process intends to determine the ability of the assessment tool to discriminate between operators with different experience level. In our study, we were able to identify significant differences of proficiency between novices and experts. Previous studies determining ICSAD construct validity have used just 2 groups. The construct validation process is successful if the tool is capable to discriminate between 2 groups.[5]

Regarding the moderate correlation, the negative sign indicates that a high score on the X variable would predict a low score on the Y variable. The absolute value of .50 suggests a fairly predictable relationship between X and Y. A correlation of -.50 would be considered a large correlation in psychological research even though it is only halfway between no relationship and a perfect relationship (our correlation was 0.467).

In this context, we think that our conclusion is valid. We have edited it in order to be concordant with these results.

24. - Line 205: Concerning the "post hoc power analysis" see above

We have improved this part.

25. - Line 209: This interpretation does not necessarily challenge the utility of a checklist. It may only demonstrate that the novices weren't novices any more. By watching the videos they obviously learned a lot about how to perform spinal anesthesia and, thus, could gather points on the checklist.

We totally agree with this point (video). However, it has been previously described that GRS are better than checklists to discriminate between operators of different expertise.[7]

26. - Line 234/235: This is not a limitation of the study, but an important result. If ICSAD fails to determine differences between novices and experts (in two similar studies) this questions ICSAD as a method and should be discussed appropriately.

We have discussed this point deeply in the discussion section.
27. - Line 250 - 254: Concerning power analysis: see general considerations

We have improved this part.

28. - Line 269-277: Concerning conclusion: see general considerations

We have improved this part.

Matthew David McEvoy, MD (Reviewer 4):

1. - There are also numerous minor typographical and grammatical errors throughout the paper, but I will not address those currently.

   Thanks for the comment. We have tried to improve the quality of the manuscript.

2. - The ICSAD assessment tool has been previously validated with other procedures in the past, including clinical performance of labor epidurals and supraclavicular blocks. Given the fact that each of these is more complex in nature than a spinal block and given that these were performed in clinical settings, why not perform this study in the clinical setting? I am currently not convinced by the explanation of the rationale in the Discussion that a partial task trainer was the proper study setting in order to add new knowledge to the literature.

   Following the argument that there is an ongoing trend to establish more objective measures of technical skill, from our point of view each procedure should have their own characterization. For example, number of movements and distance travelled by hands should be established for each procedure, in order to have validated thresholds to guide training.

   We decided to perform our study using a simulator by two reasons: First, due to ethical and safety concerns. Traditional approach for training residents implies that, if necessary, a permanent supervisor provide feedback or even intervene during the procedure affecting crude measurements and its subsequently validity and interpretation. Second, using simulators standardize the level of difficulty of the procedure for every operator.

3. - Overall, the study is well-designed and well performed (given the limitation noted above). It shows a difference between trainees who have done no spinals and those who have done a few or many. The same might be achieved by knowing a trainee's procedure count, requiring
no complex assessment. However, I do believe that the study makes the false assumption that time in training correlates with novice, intermediate, and expert technical skills. This is actually a false premise. It is generally true that skills increase with increased experience, but it is not true for all physicians, and certainly not true for technical skills where digital natives may have much better spatial reasoning and their learning curve may be significantly different than those who trained 10-20 years ago. To assess this, I would have preferred to see all participants perform 3-5 attempts. This could have assessed consistency in performance and whether there was a learning curve on the simulator (or in the clinical setting). Analysis of repeated performance over time (e.g. CUSUM) would make the conclusions more powerful. It should also be pointed out that in Dreyfus' model of skill acquisition, the levels are novice, advanced beginner, competent, proficient, and expert. Intermediate is not a level, nor should it be the goal of competency assessment. Assessing competency is the goal of most of these assessments, and it ought to be as it can determine when clinical autonomy can be granted.

We totally agree that the analysis of repeated performance over time would be valuable. The aim of this study was to demonstrate that this tool works in spinal anesthesia (construct and concurrent validity). The next step is to use this already validated tool to measure repeated performances over time, in order to build up learning curves.

We have added the explanation about the Dreyfus' model of skill acquisition in the discussion section.

4. - Building upon the prior point, given the fact that this report is not clinical and is relying upon the prior work of validation of the ICSAD tool, I would like to see an analysis of performance with a standard setting irrespective of the training level (setting of a minimum passing score). The study proposes that it distinguishes skill level, but it assumes the manner in which the groups are divided. Advancing the educational research literature will occur when competency-based assessment of success or failure (as mentioned in the Introduction) can be undertaken without reference to time in training or time past graduation, as some trainees gain skills quite quickly and some post-graduates have skills that fade. I would recommend that the authors consider assessing their data based upon the measured performance in relation to a determined standard, which in this case may need to be created (akin to Anghof or Hofstee methods).

The utility of hand motion analysis for anesthesia procedures is in validation stage. From our point of view, these tools may be complementary to visual scales. The idea to use them for competency-based assessment and to establish minimal cutoff threshold to be achieved can guide future development of this research line.
5. The discussion section seems focused on GRS superiority over TSC modalities. This discussion may represent the authors' beliefs and previous literature, but does not seem truly germane to the study, especially when only one of many measures showed any difference between intermediate and expert participants. This would actually point to the fact that such a study would be of great value if done in patients, as the best metric of block success is whether the procedure was done in a safe manner, whether the patient was satisfied, and whether the block achieved the desired clinical outcome for surgical anesthesia. Assessment of motion and time of placement may focus on proficiency and expertise, but these may also be misleading.

We have improved the discussion section with this feedback.

6. Abstract and manuscript cite different Cohen Kappa values - 0.87 v. 0.76.

We have edited this mistake. It was because we reported the kappa value for checklist in the abstract and the kappa for GRS in the manuscript.


