BJA
British Journal of Anaesthesia

Validity and reliability of an objective structured assessment tool for performance of ultrasound-guided regional anaesthesia

Journal:	*British Journal of Anaesthesia*
Manuscript ID	BJA-2017-00326-LC033.R2
Article Type:	Clinical Investigation
Date Submitted by the Author:	n/a
Complete List of Authors:	Shafqat, Atif; University of Nottingham, Anaesthesia & Critical Care Section Rafi, Mohamed; Nottingham University Hospitals NHS Trust, Department of Anaesthesia Thanawala, Vishal; Nottingham University Hospitals NHS Trust, Department of Anaesthesia Bedforth, Nigel; Nottingham University Hospitals NHS Trust, Department of Anaesthesia Hardman, Jonathan; University of Nottingham, Anaesthesia & Critical Care Section McCahon, Robert; Nottingham University Hospitals NHS Trust, Department of Anaesthesia
Mesh keywords:	Anaesthetists, Checklist, Educational assessment

Validity and reliability of an objective structured assessment tool for performance of ultrasound-guided regional anaesthesia

A. Shafqat [1], M. Rafi [2], V. Thanawala [2], N.M. Bedforth [2], J.G. Hardman [1], R.A. McCahon [2]

1. Anaesthesia & Critical Care, Division of Clinical Neuroscience, School of Medicine, University of Nottingham, UK
2. Department of Anaesthesia, Queen's Medical Centre, Nottingham University Hospitals NHS Trust, UK

Correspondence to: Dr Atif Shafqat (atif.shafqat@nhs.net)

Running head: Assessment of ultrasound-guided regional anaesthesia

ABSTRACT

Background: We aimed to examine the validity and reliability of previously developed criterion-referenced assessment checklist (AC) and global rating scale (GRS) to assess performance in ultrasound-guided regional anaesthesia (UGRA).

Methods: Twenty-one anaesthetists' single, real-time, UGRA procedures (total: 21 blocks) were assessed using 22-item *AC* and 9-item *GRS* scored on a 3-point and 5-point Likert scales respectively. We used one-way ANOVA to compare assessment scores between 3 groups (group I: ≤30 blocks in the preceding year; group II: 31–100; group III: >100). Concurrent validity was evaluated using Pearson's correlation (r). We calculated type-A intra-class correlation coefficient (ICC) using an absolute agreement definition in two-way random effects model, and inter-rater reliability (IRR) using absolute agreement between raters. The inter-item consistency was assessed by Cronbach's alpha (α).

Results: Greater UGRA experience in the preceding year was associated with better AC [F (2,18) 12.01; p <0.001] and GRS [F (2,18) 7.44; p =0.004] scores. There was strong correlation between mean AC and GRS scores [r=0.73 (p <0.001)] and strong inter-item consistency for AC (α = 0.94) and GRS (α = 0.83). The ICC (95% CI) and IRR (95% CI) for AC was 0.96 (0.95 – 0.96) and 0.91 (0.88 – 0.95) respectively and 0.93 (0.90 – 0.94) and 0.80 (0.74 – 0.86) for GRS.

Conclusions: Both assessments differentiated between individuals who have performed fewer (≤30) and many (>100) blocks in the preceding year, supporting construct validity. It also established concurrent validity and overall reliability. We recommend both tools may be used in UGRA assessment.

Keywords: Anaesthetists, Checklist, Educational assessment, Reproducibility of results, Ultrasound

INTRODUCTION

Reduced clinical opportunities during training, an increased focus on optimal patient safety, and greater public accountability have led to the need for objective assessment of procedural skills in medicine.[1,2] Assessment of expertise in medicine may be formative (developmental), or summative (pass/fail). Assessments assist practitioners towards expert practice, while protecting patients by ensuring that safe, acceptable standards of practice are maintained. Assessments must be sufficiently valid and reliable to withstand scrutiny and challenge from the learner and the patient groups; they must be credible and consistent, in order that they have value and meaning.[3,4] Following the publication of recommendations for training in ultrasound-guided regional anaesthesia (UGRA)[5], a group of 18 UGRA experts used a modified Delphi technique to develop a criterion-referenced *assessment checklist (AC)* and the *global rating scale (GRS)* to assess the technical and non-technical aspects of UGRA performance.[5,6] However, the authors stated that future work should concentrate on establishing further evidence to support the validity and reliability of these assessments. Therefore, we examined the ability of the AC and the GRS to quantify the level of expertise in UGRA in anaesthetists (construct validity). We also examined the degree of inter-rater agreement and consistency of each assessment tool, and finally, the strength of agreement between the two assessments (concurrent validity).

METHODS

We requested ethics review by the University of Nottingham Medical School Research Ethics Committee, which approved the study (Approval Reference; K09052013LT 13053 SCS Anaesthesia). Anaesthetists working at Nottingham University Hospitals NHS Trust were invited to participate in the study via email. A participant information sheet was forwarded to those who expressed an interest and written informed consent was gained in advance of any study activity. The patients of the participating anaesthetists were also given an information leaflet prior to their surgery, and their written informed consent was sought on the morning of their surgery. Each participating anaesthetist was given a participant identification number prior to commencement of the study.

This dual-site, blinded observational study was conducted concurrently at the Queen's Medical Centre and the City Hospital campuses of the Nottingham University Hospitals NHS Trust. Anaesthetists were eligible for study inclusion if they planned to perform an ultrasound-guided nerve or plexus block as part of their usual management for a patient, and the patient had agreed to take part. Exclusion criteria included anaesthetists or patients who did not wish to participate and patients who did not require UGRA. The clinical decision to perform UGRA was taken in all cases by the attending anaesthetist.

Prior to commencement of the UGRA procedure, each participant completed a self-reported questionnaire with regard to the number of ultrasound-guided nerve blocks they had completed in the preceding year. In order to minimise observer bias both the investigators were kept blinded from the completed self-reported questionnaire, which was submitted to them in a sealed envelope. Subsequently, two anaesthetist investigators (AS & MR) observed the participants together and used AC and GRS to independently assess UGRA performance by participants during routine operating lists. Assessment occurred in real-time during performance of a single UGRA procedure by each participant, and began with the initial preparation and set-up of equipment and ended at completion of the procedure. The two investigators completed the assessments simultaneously and did not influence the clinical practice of the participants in any way.

AS and MR had been trained to use both assessment tools (AC and the GRS) before study commencement. This involved a week of practice assessments sessions (5 half-days) with facilitated debriefing from the research team so that both assessors were familiar with the assessment tools, and that they had a shared understanding of UGRA performance. In brief, the AC comprises of 22 items scored on a 3-point Likert scale (not performed [0], poorly performed [1], well performed [2]) (appendix 1), whereas the GRS consists of nine categories scored on a 5-point Likert scale with descriptive anchors of performance to assist scoring (appendix 2).[6] One of the categories of GRS that is the item "*overall performance*" was excluded from the calculation of GRS score. In addition to that, we did not record a "*Pass / Fail*" assessment.

Statistics

In line with previous studies, we estimated that we would need to recruit between 20 and 40 participant anaesthetists.[7-11] For the purpose of analysis, we arbitrarily allocated all the participants to one of the three groups, based on the self-reported questionnaire with regard to the number of ultrasound-guided nerve blocks they had completed in the preceding year (group I: ≤30; group II: 31–100; group III: >100).

Statistical analysis used STATA/IC version 10.0 (StataCorp, Texas, USA).

Normality of data was assessed by histogram and the Shapiro–Wilk and Skewness/Kurtosis tests. To test whether higher total assessment scores were associated with a greater number of ultrasound-guided nerve blocks in the preceding year, we used a one-way ANOVA to compare AC and GRS scores for groups I–III. Where a significant difference was identified, we performed appropriate *post hoc* comparisons with a Bonferroni adjustment to adjust for multiple comparisons.

We performed an exploratory analysis of relationships between the values of AC score, GRS score, response to GRS item *"overall performance"* and number of blocks in the preceding year by calculating the Spearman correlation coefficient ρ (rho). Similarly, we evaluated the concurrent validity of the assessment tools by calculating the Pearson correlation coefficient (r). In all analyses, we used a two-tailed p-value less than 0.05 to indicate statistical significance.

To assess inter-rater agreement, we calculated the type-A intra-class correlation coefficient (ICC) using an absolute agreement definition, in a two-way random effects model. We also calculated inter-rater reliability (IRR) using the proportion of absolute agreement between raters:[9,12]

$$\mathrm{IRR} = \frac{Observation\ event\ agreements}{Total\ number\ of\ observations}$$

We calculated Cronbach's alpha (α) coefficient to assess inter-item consistency within each assessment tool.

RESULTS

Twenty-one anaesthetists enrolled in the study; no subject dropped out. Each participant performed a single UGRA procedure (total blocks: 21), which was observed in real-time during routine operating lists. The median [IQR (Range)] number of ultrasound-guided nerve blocks completed by all the participating anaesthetists in the preceding year was 90 [30–160 (5–600)]. Seven participants were allocated to group I (≤ 30 blocks in preceding year), 6 to group II (31 – 100 blocks), and 8 to group III (> 100 blocks). The mean (SD) number of ultrasound-guided nerve blocks completed by the participants in the preceding year was; group I - 22.28 (10.56), group II - 78.33 (23.16) and median [IQR (Range)] for group III - 200 [155–225 (109–600)] respectively. The type of ultrasound-guided blocks observed are summarised in Table 1. Score data from the AC and the GRS were considered to be normally distributed.

Assessment Checklist (AC):

The maximum achievable score on the AC is 44. The mean (SD) AC score was 32.85 (5.22) for all participants. One-way ANOVA revealed an overall statistically significant difference in the AC scores between groups I, II & III [F (2,18) 12.01; p <0.001], with a mean rank AC score of 28.28 (group I), 32.16 (group II) and 37.37 (group III) (Fig. 1). *Post hoc* comparison of the group means revealed a statistically significant difference between groups I and III (p <0.001) and between groups II and III (p =0.04). There was no significant difference between groups I and II.

There was a strong correlation (ρ 0.704; p <0.001) between AC scores and the total number of UGRA nerve blocks performed in the preceding year (i.e. a larger number of UGRA nerve blocks in the preceding year was associated with better AC scores) (Fig. 2).

Global Rating Scale (GRS):

The maximum achievable score in the GRS is 40. The mean (SD) GRS score was 30.09 (7.15) for all participants. One-way ANOVA revealed an overall statistically significant difference in the GRS scores between groups I, II & III [F (2,18) 7.44; p =0.004] with a mean rank GRS score of 24.57 (group I), 29.16 (group II) and 35.62 (group III) (Fig. 3). *Post hoc* comparison of the group means revealed a statistically significant difference between groups I and III (p =0.004). There was no significant difference between groups I and II and between groups II and III.

There was a strong correlation (ρ 0.712; p <0.001) between GRS scores and the total number of UGRA nerve blocks performed in the preceding year (i.e. a larger number of UGRA nerve blocks in the preceding year was associated with better GRS scores) (Fig. 4).

Reliability of the assessment tool:

There was a strong correlation (ρ 0.90; p <0.001) between GRS score and response to GRS item *"overall performance"* and a strong correlation between AC and GRS scores (r 0.73; p <0.001).

The ICC and the IRR are summarized in Table 2. Cronbach's alpha coefficient (standard error of measurement) for the AC and the GRS were 0.94 (3.41%) and 0.83 (10.25%) respectively, indicating strong inter-item consistency.

DISCUSSION

In this study of UGRA performance, we have demonstrated that both the AC and the GRS may be used to differentiate between individuals who have performed fewer (≤ 30) and many (> 100) nerve blocks in the preceding year; this finding is consistent with the findings of the recent study of Ahmed and colleagues.[12] However, neither assessment tool was sufficiently sensitive to identify the anaesthetists who had performed an intermediate number of nerve blocks (31–100) in the preceding year. This could represent a type-2 error, rather than lack of sensitivity in the assessment tools, although our sample size is in line with similar and recent studies.[9-11] Our study results demonstrate that both the AC and the GRS have appropriate construct validity. Furthermore, there is strong correlation between the assessment scores and the number of nerve blocks performed recently. In conjunction with the strong correlation between the results of the AC and the GRS, these results demonstrate appropriate concurrent validity. With regard to reliability of the assessment tools, our results indicate strong inter-rater agreement and inter-item consistency.

Mastery of a complex task requires effective feedback to guide deliberate practice.[13] Our results indicate that the AC and the GRS are sufficiently reliable formative assessment tools to provide effective feedback. In particular, both assessments provide useful anchors upon which feedback to learners can be based. Whether these assessment tools could be effective when used summatively is less clear. Key considerations would be who, when, how, and to what end these assessments are to be applied. For moderate stakes assessments, e.g. major end-of-course tests, it is held that reliability coefficients should be at least 0.8–0.89;[4] we have demonstrated reliability coefficients greater than 0.8 for both assessments in all analyses except one, thus replicating our previous analysis of the GRS.[14] On this basis, it would be reasonable for the AC and the GRS to be used in pass/fail assessments at the end of a unit of training in UGRA; this could equally be applied to doctors in training or consultant staff who are new to UGRA or refreshing their skills. At the other end of the expertise spectrum, we are reluctant to suggest that either assessment be used to inform very high stakes decisions, e.g. to grant licensure or accreditation. At the very least, any assessment outcome would need to be considered in the context of a practitioner's clinical outcomes, before a truly informed, valid and reliable decision could be made.

Wong and colleagues,[10] reported UGRA performance data in 13 trainee anaesthetists, using a modified checklist and the GRS; this study was published soon after recruitment to our study was completed. In broad terms, our findings are in agreement in that we found that both assessment tools are reliable and possess construct and concurrent validity; in addition, we have demonstrated improved inter-rater agreement, with much less variability in assessment, i.e. greater precision. Unlike Wong's study, where six multi-centre raters were used; we had two raters based at the same centre, one of whom had previously assessed UGRA performance of 60 recruits using the GRS and *composite error score*.[14] In that study, we reported ICC of 0.91 and 0.97 for the GRS and the *composite error score*, respectively. Caution is required in interpreting the high ICC, as it is increased by greater between-subject variance, and is thus affected by subjects' expertise. We studied a broader spectrum of expertise, which may have contributed to the greater ICC in our study. To account for this statistical effect, we presented sub-group analyses for each expertise level, and we found that the strong inter-rater agreement remained. For each participant both the raters scored the AC and the GRS consecutively. The scoring for one tool could have influenced the scores for the other tool, which could be a limitation and partially explain the high level of agreement between scores for the two different tools.

The use of non-binary checklists or Likert-scales for assessment in UGRA, and analysis of reliability using ICC could be considered weaknesses in our study.[9,12] Gallagher and colleagues[9] argue that high-stakes assessment should be made using binary assessments with high IRR, e.g. percentage agreement, as this approach is less ambiguous, and is more transparent and defensible. In contrast, the AC and the GRS introduce subjectivity in the assessment of performance, thus limiting reliability and defensibility. While we have demonstrated strong agreement for both assessments, it is evident that there is less consistent agreement using the GRS, which has more rating choices in its 5-point Likert-scale. The real-world significance of this is debatable, as both rating tools performed adequately (and similarly) in differentiating expertise levels in anaesthetists. A recent systematic review (considering 45 studies) concluded that the GRS are better able to discriminate expert performance, and are more reliable than their dichotomous rivals.[15] As such, we support and recommend previous assertion that the combination of validated checklists and the GRS by trained assessors is the current gold standard for assessment in UGRA.[2] With regard to the issues of ICC, previous authors have challenged the wisdom of using this analysis as a measure of inter-rater agreement, arguing that it measures the association between raters' decisions rather than their absolute agreement.[9] To mitigate this, we calculated the type-A ICC using an absolute agreement definition, in a two-way random effects model, measuring *absolute agreement* and the *correlation* between raters' score differences. For the sake of comparison, we also calculated the IRR or *absolute agreement*. The calculated coefficients were very high, but consistently lower than the corresponding ICC, and with greater associated variability. Nevertheless, our findings were very similar to those of Ahmed and colleagues,[12] who are proponents of IRR and critics of ICC for the evaluation of assessment reliability. While absolute rater agreement is appealing in its simplicity, it is not necessary for high stakes assessment; indeed it fails to account for chance agreement.[4] Instead, what counts most is consistency of assessment between raters, which is achieved through rater training and regular quality assurance – small differences in item ratings do not render an assessment tool unreliable.

We must acknowledge that one rater (MR) worked at both campuses of Nottingham University Hospitals, and personally knew most of the study participants; therefore, this could risk adding rater bias. However, the other rater (AS) had never worked in either department, and was oblivious to the study participants' grade or their previous UGRA practice. The strong inter-rater agreement indicates that any potential rater bias had minimal impact.

Another key limitation in the interpretation of our findings is generalizability (external validity). This problem is likely common to all assessments in UGRA, and in particular, the wording of performance descriptors may result in varying assessment scores for the same individual UGRA block performance, reflecting variable interpretation of performance descriptors by local raters. This issue is not insurmountable; a decision as to how the assessments are used, i.e. high/moderate/low stakes assessment is therefore required. Centralisation of future high stakes UGRA assessment would ensure a shared frame of reference upon which expert raters can base their decisions and ensure a regular cycle of quality assurance.

Lastly, there is the question of whether *block-specific* or *block-generic* assessments should be used. There is no agreement on a validated technique or tool to evaluate learners' performance which would permit assessment between different organizations.[16] In the past, block-specific assessments had been developed for obstetric epidural analgesia,[17] spinal anaesthesia,[18] interscalene brachial plexus blocks,[7] ultrasound-guided supraclavicular blocks,[8] and ultrasound-guided axillary brachial plexus blocks.[19] Block-generic assessments was initially formulated by Cheung and colleagues[6] for all peripheral UGRA procedures, which was validated by Wong and colleagues[10] and Burckett-St Laurent and co-workers.[20] Recently, Chuan and others[11] have designed block-generic assessments, which could be used for any type of regional anaesthesia procedures. The advantage of block-generic over block-specific assessments is their wider capability to assess different types of regional anaesthesia procedures. In all these assessments, unlike GRS, the checklists used were not similar in design, demonstrating inconsistent validity and reliability.[21] Therefore, it requires further research, and if consistent validity and reliability is successfully established, then there is a promising chance that a single homogenous checklist layout could be formulated.

In summary, we believe that we have presented a robust argument for the use of the AC and the GRS for low/moderate stakes assessment in UGRA. In terms of their utility, both assessments appear reliable and valid. We do not yet know if these assessments and others like them are acceptable to learners and whether they have any educational impact on expertise acquisition. Future research should examine whether these competency-based assessment tools result in improved expertise gain.

AUTHOR CONTRIBUTIONS:

A.S.	Study design, participant recruitment, data collection, data analysis, critical revision and writing up the first & final draft of the paper.
M.R.	Participant recruitment, data collection, review of draft paper.
V.T.	Participant recruitment, data collection, review of draft paper.
N.M.B.	Study design, critical revision of draft papers.
J.G.H.	Study design, critical revision of draft papers.
R.A.M.	Study concept, study design, data analysis, critical revision and writing up of the final draft of the paper.

ACKNOWLEDGEMENTS

We would like to thank the Anaesthetic staff and patients of Nottingham University Hospitals NHS Trust, without their co-operation this study would not have been possible.

DECLARATION OF INTERESTS

A.S.	Nil to declare
M.R.	Nil to declare
V.T.	Nil to declare
N.M.B.	Nil to declare
J.G.H.	I receive research funding from industry, charities and research councils. I receive payment for medicolegal work from solicitors, coroners and the police. I am associate editor in chief and a member of the editorial board of the British Journal of Anaesthesia.
R.A.M.	Nil to declare

FUNDING

This work was supported by departmental funds.

REFERENCES

1. Epstein RM: Assessment in medical education. N Engl J Med 2007; 356: 387-96

2. Bould MD, Crabtree NA, Naik VN: Assessment of procedural skills in anaesthesia. Br J Anaesth 2009; 103: 472-83

3. Downing SM: Validity: on meaningful interpretation of assessment data. Med Educ 2003; 37: 830-7

4. Downing SM: Reliability: on the reproducibility of assessment data. Med Educ 2004; 38: 1006-12

5. Sites BD, Chan VW, Neal JM, Weller R, Grau T, Koscielniak-Nielsen ZJ, Ivani G: The American Society of Regional Anesthesia and Pain Medicine and the European Society of Regional Anaesthesia and Pain Therapy joint committee recommendations for education and training in ultrasound-guided regional anesthesia. Reg Anesth Pain Med 2010; 35: S74-S80.

6. Cheung JJ, Chen EW, Darani R, McCartney CJ, Dubrowski A, Awad IT: The creation of an objective assessment tool for ultrasound-guided regional anesthesia using the Delphi method. Reg Anesth Pain Med 2012; 37: 329-33

7. Naik VN, Perlas A, Chandra DB, Chung DY, Chan VW: An assessment tool for brachial plexus regional anesthesia performance: establishing construct validity and reliability. Reg Anesth Pain Med 2007; 32: 41-5

8. Chin KJ, Tse C, Chan V, Tan JS, Lupu CM, Hayter M: Hand motion analysis using the imperial college surgical assessment device: validation of a novel and objective performance measure in ultrasound-guided peripheral nerve blockade. Reg Anesth Pain Med 2011; 36: 213-9

9. Gallagher AG, O'Sullivan GC, Leonard G, Bunting BP, McGlade KJ: Objective structured assessment of technical skills and checklist scales reliability compared for high stakes assessments. ANZ J Surg 2014; 84: 568-73

10. Wong DM, Watson MJ, Kluger R, Chuan A, Herrick MD, Ng I, Castanelli DJ, Lin LC, Lansdown A, Barrington MJ: Evaluation of a task-specific checklist and global rating scale for ultrasound-guided regional anesthesia. Reg Anesth Pain Med 2014; 39: 399-408

11. Chuan A, Graham PL, Wong DM, Barrington MJ, Auyong DB, Cameron AJ, Lim YC, Pope L, Germanoska B, Forrest K, Royse CF: Design and validation of the Regional Anaesthesia Procedural Skills Assessment Tool. Anaesthesia 2015; 70: 1401-11

12. Ahmed OM, O'Donnell BD, Gallagher AG, Breslin DS, Nix CM, Shorten GD: Construct validity of a novel assessment tool for ultrasound-guided axillary brachial plexus block. Anaesthesia 2016; 71: 1324-1331

13. Ericsson KA, Krampe RT, Tesch-Romer C: The role of deliberate practice in the acquisition of expert performance. Psychological Review 1993; 100: 363-406

14. Shafqat A, Ferguson E, Thanawala V, Bedforth NM, Hardman JG, McCahon RA: Visuospatial Ability as a Predictor of Novice Performance in Ultrasound-guided Regional Anesthesia. Anesthesiology 2015; 123: 1188-97

15. Ilgen JS, Ma IW, Hatala R, Cook DA: A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ 2015; 49: 161-73

16. Nix CM, Margarido CB, Awad IT, Avila A, Cheung JJ, Dubrowski A, McCartney CJ: A scoping review of the evidence for teaching ultrasound-guided regional anesthesia. Reg Anesth Pain Med 2013; 38: 471-80

17. Friedman Z, Katznelson R, Devito I, Siddiqui M, Chan V: Objective assessment of manual skills and proficiency in performing epidural anesthesia--video-assisted validation. Reg Anesth Pain Med 2006; 31: 304-10

18. Breen D, Bogar L, Heigl P, Rittberger J, Shorten GD: Validation of a clinical assessment tool for spinal anaesthesia. Acta Anaesthesiol Scand 2011; 55: 653-7

19. Sultan SF, Iohom G, Saunders J, Shorten G: A clinical assessment tool for ultrasound-guided axillary brachial plexus block. Acta Anaesthesiol Scand 2012; 56: 616-23

20. Burckett-St Laurent DA, Niazi AU, Cunningham MS, Jaeger M, Abbas S, McVicar J, Chan VW: A valid and reliable assessment tool for remote simulation-based ultrasound-guided regional anesthesia. Reg Anesth Pain Med 2014; 39: 496-501

21. Chuan A, Wan A, Royse CF, Forrest K: Competency-Based Assessment Tools for Regional Anaesthesia: A Narrative Review. Br J Anaesth 2018; 120(2): 264-273

Table 1. Type of UGRA blocks observed

Type of UGRA block	**N (%)**
Ankle	1 (4.7)
Axillary	2 (9.5)
Femoral	9 (42.8)
Interscalene	7 (33.3)
Sciatic (popliteal fossa)	1 (4.7)
Rectus sheath catheter	1 (4.7)
Total number of blocks observed	21

Table 2. Intra-class correlation (ICC) and Inter-rater reliability (IRR) for Assessment Checklist (AC) and Global Rating Scale (GRS) score. CI, confidence interval.

	AC		**GRS**	
	ICC (95% CI)	*IRR* (95% CI)	*ICC* (95% CI)	*IRR* (95% CI)
Sample (n=21)	0.96 (0.95–0.96)	0.91 (0.88–0.95)	0.93 (0.90–0.94)	0.80 (0.74–0.86)
Group I (n=7) ≤30 blocks	0.95 (0.93–0.97)	0.90 (0.84–0.96)	0.87 (0.79–0.92)	0.75 (0.59–0.90)
Group II (n=6) 31–100 blocks	0.96 (0.95–0.96)	0.94 (0.90–0.98)	0.91 (0.84–0.95)	0.80 (0.68–0.91)
Group III (n=8) >100 blocks	0.94 (0.92–0.95)	0.90 (0.81–0.99)	0.86 (0.78–0.92)	0.85 (0.74–0.96)

FIGURE LEGENDS

Figure 1: Mean (SD) assessment checklist (AC) scores for groups I – III. SD, standard deviation. Number of blocks in previous year: Group I (≤30 blocks); group II (31–100 blocks); group III (>100 blocks).

Figure 2: Relationship between assessment checklist (AC) scores *vs.* total number of UGRA nerve blocks performed in the previous year.

Figure 3: Mean (SD) global rating scale (GRS) scores for groups I – III. SD, standard deviation. Number of blocks in previous year: Group I (≤30 blocks); group II (31–100 blocks); group III (>100 blocks).

Figure 4: Relationship between global rating scale (GRS) scores *vs.* total number of UGRA nerve blocks performed in the previous year.

APPENDIX 1

Assessment Checklist Data Collection Form[6]

Participant's Study Ref:

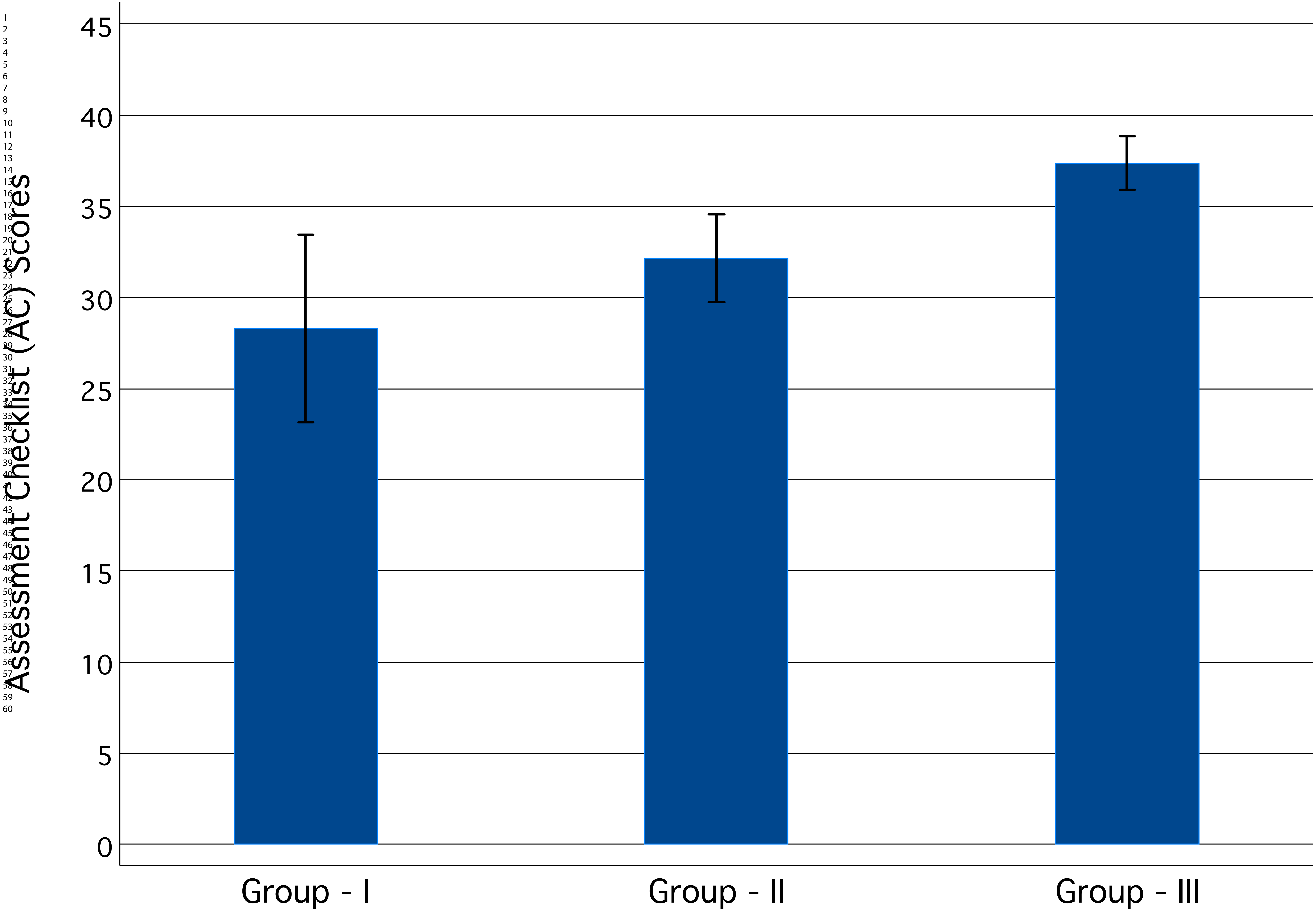
Tasks	**Not Performed 0**	**Performed Poorly 1**	**Performed Well 2**
1) Proper positioning of patient			
2) Correct placement of ultrasound machine relative to patient to allow easy visualization of both			
3) Choice of correct transducer			
4) Correct depth, gain and focal zone choices			
5) Holds the probe appropriately (3 fingers holding the probe and 1 finger touching the patient)			
6) Knowledge or confirmation of screen orientation (ie, which side of probe corresponds to which side of screen)			
7) Scanning of anatomy and proper identification of target			
8) Use of Doppler to rule out vascular structures (if applicable)			
9) Appropriate needle alignment			
10) Maintenance of needle tip image during advancement of needle			
11) Efficiency of regaining needle tip position image (PART Manoeuvre)			
12) Recognition of proper nerve stimulation at appropriate levels (if nerve stimulation used)			
13) Ensure that current is not <0.2 mA (if nerve stimulation used)			
14) Ask for initial aspiration to rule out intravascular injection			
15) Visualization of needle tip before injection			
16) Ask for 1- to 2-mL initial injection to rule out intraneural and intravascular injection			
17) Ask patient or at least look for signs of pain/discomfort			
18) Ask for proper aspiration every 5-mL increments injection			
19) Recognition of proper needle tip position			
20) Perform appropriate needle tip adjustments			
21) Assessment of ease of injection (high pressure)			
22) Recognition of correct local anaesthetic spread in relation to nerve			

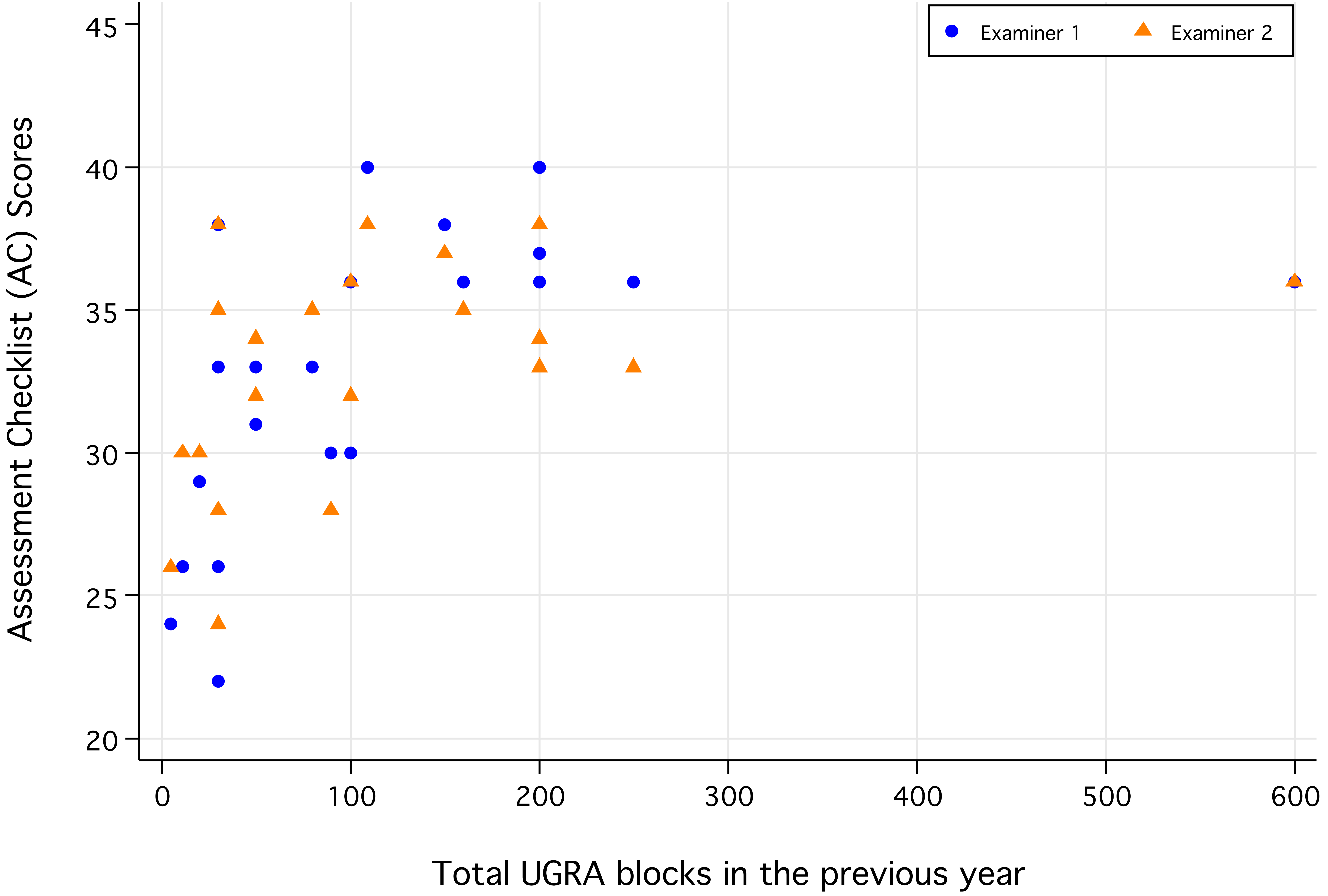
APPENDIX 2

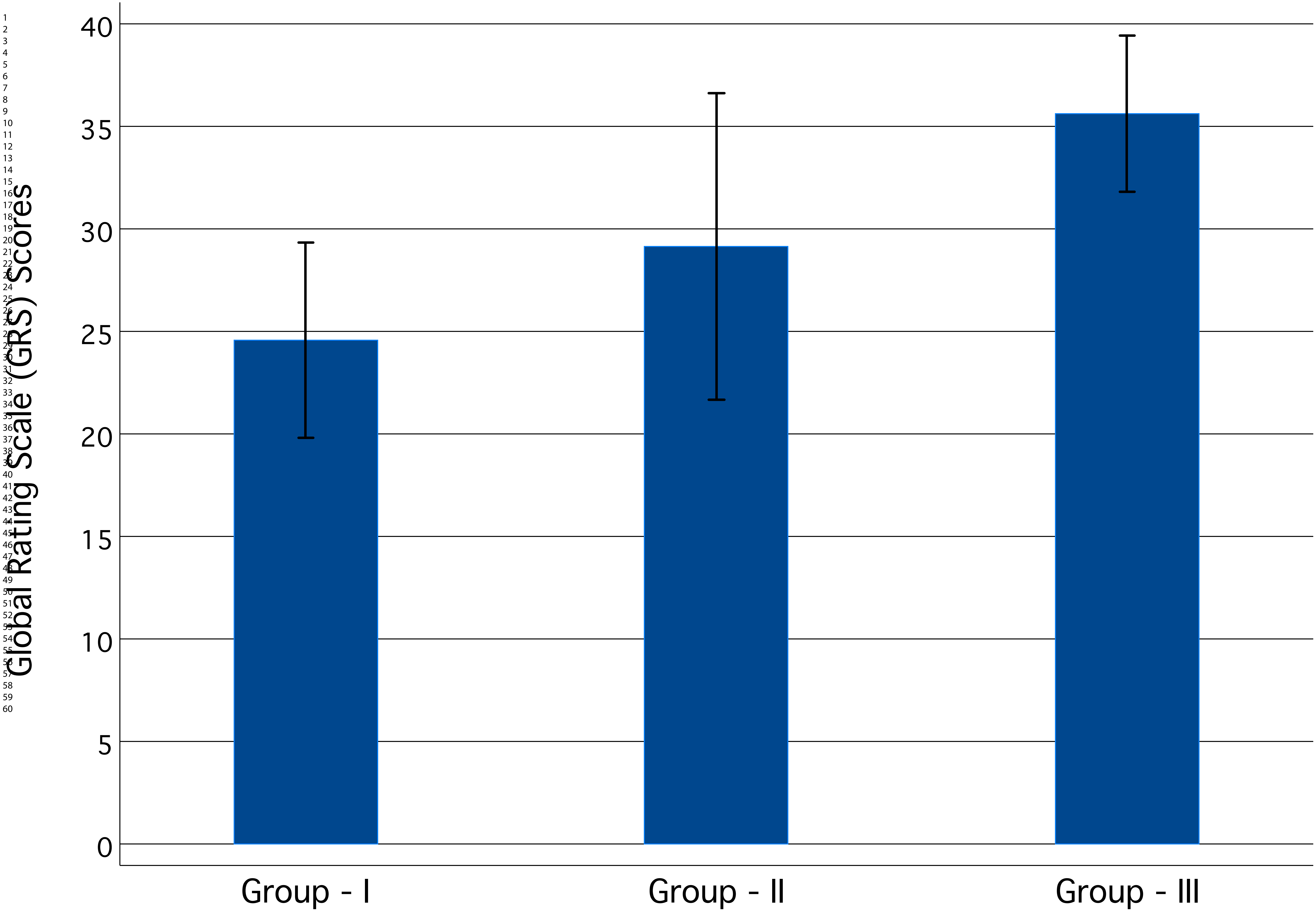
Global Rating Scale Data Collection Form[6]

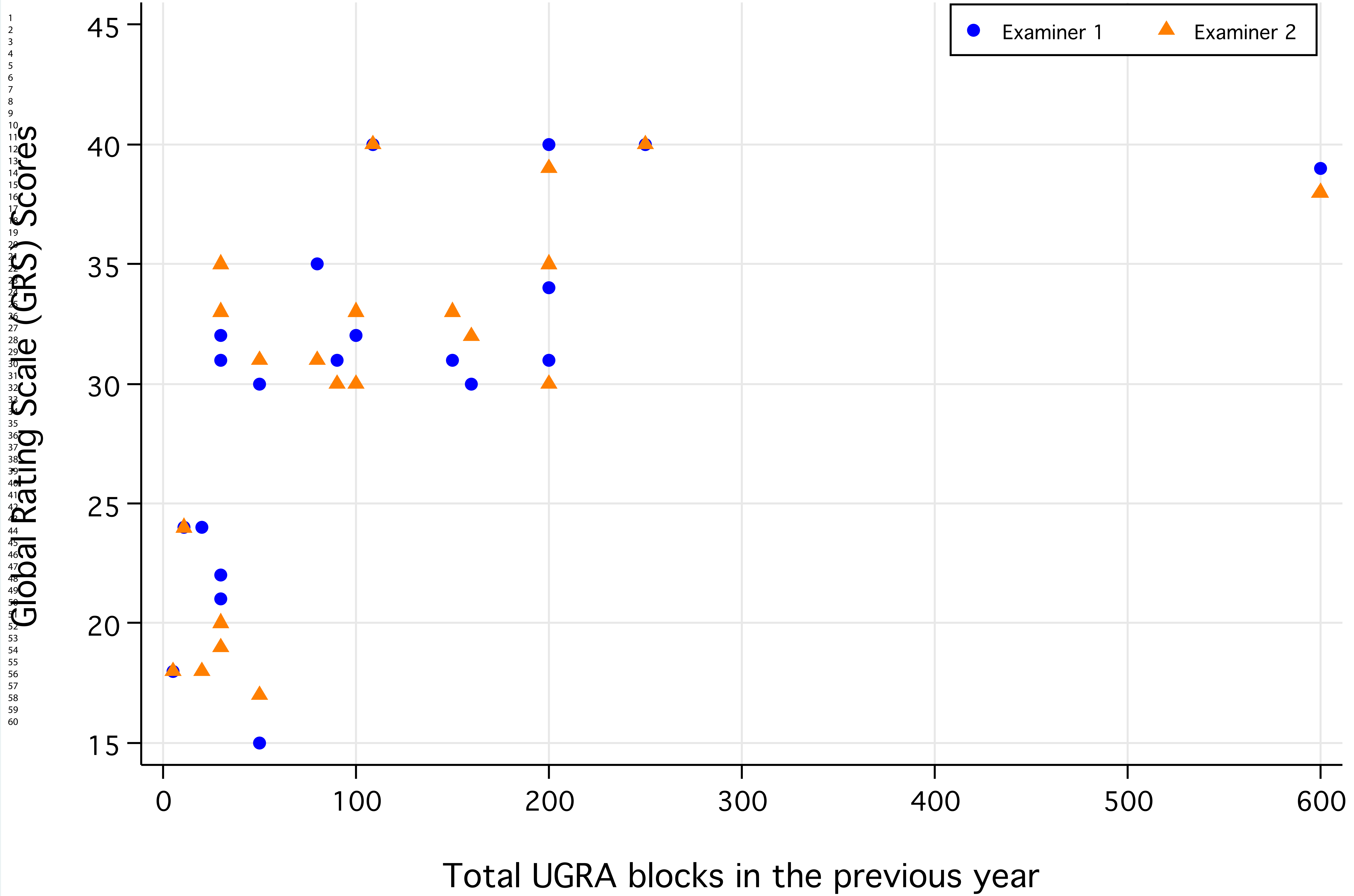
Participant's Study Ref:

	Score 1	2	3	4	5
Preparation for procedure (eg, monitors, IV access, U/S machine)	Did not organize equipment well. Has to stop procedure frequently to prepare them. [ ]	[ ]	Equipment generally organized. Occasionally has to stop and prepare them. [ ]	[ ]	All equipment neatly organized, prepared and ready for use. [ ]
Patient Interaction	Little to no rapport established; Patient is unaware of procedures. No sedation is provided. [ ]	[ ]	Rapport is generally established; patient is aware and informed of most procedures. Patient anxiety is alleviated adequately with a sedative(s). [ ]	[ ]	Strong rapport is established and maintained throughout procedure. Patient is well informed of all relevant procedures. Patient anxiety alleviated through sedation and verbal comforting. [ ]
Asepsis (eg, use of sterile gloves, proper patient draping, probe sterility, cleansing of skin before infilteration, use of op site)	Practice of proper aseptic technique not generally apparent. Many errors in aseptic technique made throughout procedure. [ ]	[ ]	Generally practices proper aseptic technique. Occasional errors in aseptic technique made during procedure. [ ]	[ ]	Excellently demonstrates proper aseptic technique. Few or no errors in aseptic technique made during procedure. [ ]
Respect for Tissue	Frequently uses unnecessary force on tissue or causes damage. [ ]	[ ]	Carefully handles tissue but occasionally causes unintentional damage. [ ]	[ ]	Consistently handles tissues appropriately with minimal damage. [ ]
Time and motion	Many unnecessary movements. [ ]	[ ]	Efficient time/motion but some unnecessary movements. [ ]	[ ]	Clear economy of movements. Maximum efficiency. [ ]
Instrument handling	Repeatedly makes tentative and awkward movements. [ ]	[ ]	Competent with instruments but occasionally makes awkward or stiff movements. [ ]	[ ]	Fluid movements with instruments and no awkwardness. [ ]
Flow of procedure	Frequently stops procedure and seems unsure of next move. [ ]	[ ]	Demonstrates some forward planning with reasonable progression of procedure. [ ]	[ ]	Obviously planned course of procedure, with effortless flow from one move to the next. [ ]
Knowledge of procedure	Deficient knowledge [ ]	[ ]	Knows all important steps of procedure. [ ]	[ ]	Demonstrates familarity with all aspects of the procedure. [ ]
Overall performance	Very poor		Competent		Clearly superior
Overall, should the candidate; Pass / Fail	[ ]	[ ]	[ ]	[ ]	[ ]


Assessment Checklist (AC) Scores
45
40
35
30
25
20
15
10
5
0
Group - I
Group - II
Group - III


Examiner 1
Examiner 2
Assessment Checklist (AC) Scores
45
40
35
30
25
20
0
100
200
300
400
500
600
Total UGRA blocks in the previous year


Global Rating Scale (GRS) Scores
40
35
30
25
20
15
10
5
0
Group - I
Group - II
Group - III


Examiner 1
Examiner 2
Global Rating Scale (GRS) Scores
45
40
35
30
25
20
15
0
100
200
300
400
500
600
Total UGRA blocks in the previous year