Which Global Rating Scale? A comparison of the ASSET, BAKSSS and IGARS for the assessment of simulated arthroscopic skills.

Abstract:

Background**:* With the move to competency-based models of surgical training a number of assessment methods have been developed. Of these Global Rating Scales (GRS) have emerged as popular tools, and several exist specific to the assessment of arthroscopic skills. Our aim was determine which of a group of commonly used Global Rating Scores demonstrated superiority in the assessment of simulated arthroscopic skills. ***Methods**:* 63 individuals of varying surgical experience performed a number of arthroscopic tasks on a virtual reality simulator (VirtaMed ArthroS™). Performance was blindly assessed by 2 observers using 3 commonly used GRSs used to assess simulated skills. Performance was also assessed by validated objective motion analysis. ***Results**:* All the GRSs demonstrated construct validity with significant differences between each skill level and each arthroscopic task (*p<0.002*, Mann-Whitney U test). Inter-rater reliability was excellent for each GRS (overall Cronbach $\alpha = 0.907$). Correlations of GRS ratings to motion analysis was high and strong for each GRS when correlated to time taken (Spearman's rank, r_s>-.75, p<0.001), while correlation with 'total path length' was significant and moderately strong (r_s>-.64, p<0.001). ***Conclusions: No one GRS demonstrated superiority as an assessment tool. ***Clinical relevance**:* For these commonly used arthroscopic GRSs, none were particularly superior and any one score could therefore be used. Agreement on using a single score seems sensible, and it would seem unnecessary to develop further scales with the same domains for these same purposes.

Introduction

The classic apprenticeship model of surgical training is no longer viable. Recent years have seen the introduction of enforced limits on working hours, aiming to improve patient safety and doctor wellbeing.[1–4] However, such measures have had a reportedly negative impact on surgical training.[5,6] No longer is a single master responsible for the training of his apprentice, instead a number of trainers are involved over relatively short periods. Whilst this improves the breadth of training for the trainee, it becomes more difficult for the trainers to assess a trainee's progress and skill acquisition. The reduction in working hours has also led to reduced opportunities for training in the operating room and clinics, vital components of surgical training.[7] To meet these challenges training programmes are adopting competency-based models, in which objective skill assessments are required for training progression.[8,9]

To address the transition from apprenticeship-based to competency-based training, methods for the objective assessment of surgical skill have been developed. Common examples are motion analysis, task-specific check lists and Global Rating Scales (GRSs).[10–14] Motion analysis has been validated as an assessment tool in determining skill level, however there are associated equipment costs and practical limitations.[10] Task-specific check lists are the least powerful of the tools, in that they are applicable to a single task, and have demonstrated early ceiling effects.[15,16] Finally, GRSs have been developed as generic tools that assess performance

across several skill domains, to generate an overall performance rating. GRSs have been utilized in multiple surgical specialties in both simulated and clinical settings.[17] GRSs designed for use in arthroscopic surgery have demonstrated the ability to assess both subject skill levels and learning curves. [15,18] GRSs are an attractive and feasible assessment tool given the lack of additional equipment required, usability in the operating room, applicability to multiple tasks and freedom on assessment timings. **(CME1)**

A number of arthroscopy GRSs have been developed at different institutions, and their psychometric properties have been validated for use in the skill laboratory setting.[19–21] They aim to be generic to simulator format (cadaveric vs. simulated vs. benchtop) and simulated joint. To date each of these GRSs have been assessed in isolation from one another, and it is not clear whether there are any advantages to any one over the others. The aim of this study was to compare the reliability and construct validity of three previously validated GRSs in order to determine which might be superior.

Methods

Setting & Subjects

This study was undertaken in the surgical skills laboratory of a university teaching hospital. 63 participants volunteered; 31 individuals were classified as 'Novices' with no independent arthroscopic experience; 21 individuals in current orthopaedic training programmes and with less than 100 independent arthroscopies were classified as 'Trainees', and 11 individuals with more than 100 independent arthroscopies were classified as 'Expert'**.** A cut-off of 100 arthroscopies was selected on the basis that this is in excess of the minimum requirements in our country before the end of training, and that for both simple and complex simulated arthroscopic tasks in knee and shoulder models requires >100 performances (unpublished data). Concealed randomization assigned individuals to a simulated knee or shoulder joint for assessment. This resulted in 16 novices, 11 trainees and 6 expert individuals assigned to the knee model, and 15 novices, 10 trainees and 5 expert individuals assigned to the shoulder model. With Institutional review board approval (MSD-IDREC-C1-2014-152) each participant gave informed consent for participation. No participants had experience with the virtual reality (VR) arthroscopic simulator used.

Tasks

All simulated tasks were performed on the VirtaMed ArthroS™ VR simulator (VirtaMed AG, Zurich, Switzerland), a VR arthroscopic simulator with passive haptic feedback.[22] The standard objective simulator output dexterity metrics were recorded (total time, camera path length, instrument path length). Video footage of each arthroscopic task was recorded for review by 2 blinded assessors. Both assessors were orthopaedic trainees who have been trained in the use of GRSs by senior orthopaedic surgeons, with one assessor having used GRSs extensively in prior studies. We selected 2 assessors as sufficient on the basis that two of the GRSs utilised in the study were developed with 2 assessors, and that prior work by our group has found that additional assessors does not add to the results. Standardized arthroscopic tasks are a feature of the simulator, in which individual steps are presented to the participant, and these need to be completed to progress through the simulation.

Each participant performed three tasks on the simulated knee or shoulder joint, with the tasks being performed in succession. The arthroscopic knee tasks consisted of a diagnostic knee arthroscopy (arthroscope only), a basic triangulation task with fixed targets (arthroscope and hook) and a medial menisectomy (arthroscope, punch, shaver and hook). The arthroscopic shoulder tasks consisted of a diagnostic shoulder arthroscopy (arthroscope), a basic triangulation task with fixed targets (arthroscope and hook), and an advanced triangulation task with free-floating targets (arthroscope and hook). All tasks were performed with an investigator present, on-screen guidance disabled, standardized instrument portals and no time limit.

Global Rating Scales

Three validated and commonly used GRSs were selected for comparison, each developed to assess global arthroscopic skills. The Arthroscopic Surgical Skill Evaluation Tool (ASSET) was developed by Koehler *et al* as an assessment tool for diagnostic arthroscopy of cadaveric knee specimens.[19] The Basic Arthroscopic Knee Skill Scoring System (BAKSSS) developed by Insel *et al* consisted of both a task-specific checklist and GRS for arthroscopic tasks of the knee. It has since been validated utilizing only the GRS component.[20] Finally, the Imperial Global Arthroscopy Rating Scale (IGARS) was developed by Bayona *et al* as an assessment tool for both real and simulated arthroscopic tasks.[21] Each of these GRSs allows assessors to rate performance across several domains, using 5 point Likert scales with anchors at 1, 3 and 5 points.

Given the use of these scales in simulated arthroscopic tasks, certain domains were not applicable and so some modifications were required to each GRS such that they were appropriate for the tasks in hand. However, the experiments would then determine if the GRSs retained their validity as assessment measures in this simulated scenario. We have previously demonstrated this with the BAKSSS GRS, which retained its validity.[18] Similarly, the need for additional equipment in the IGARS GRS reduces its ease of use in real world settings, and thus this domain was also omitted. A summary of the changes to each GRS is provided below.

ASSET - For the diagnostic task requiring only the use of the arthroscope, domains 4 (instrument dexterity) and 5 (bimanual dexterity) were excluded. No domains were excluded in the triangulation or menisectomy tasks.

BAKSSS - Domains 1 (dissection), 4 (bimanual dexterity), 6 (knowledge of instruments), 8 (knowledge of specific procedure) and 9 (autonomy) were excluded in the diagnostic task. Domains 1, 6, 8, and 9 were excluded in the triangulation and menisectomy tasks. Domain 10 (quality) was determined by the percentage of chondral damage recorded by the simulator (<20% = 5, 21-40% = 4, 41-60% = 3, 61-80% = 2, >80% = 1).

IGARS - Domains 4 (proficiency with instruments excluding arthroscope), 6 (bimanual dexterity and triangulation) and 9 (knowledge of specific procedure/tasks to be performed) were excluded from the diagnostic task. Domain 9 was excluded in the triangulation and menisectomy tasks.

Statistical Analysis

The Shapiro-Wilk test showed the GRS scores of participants were not all parametrically distributed, thus for statistical robustness non-parametric tests were applied throughout.

The Kruskal-Wallis test was applied to each task to determine if significant differences existed between groups. Where this was the case the Mann-Whitney-U test was used to determine which groups were significantly different from one another. Similar methods were used to assess the dexterity metrics (camera path length and instrument path length) recorded by the simulator. The Spearman rank correlation was used to assess correlation between each GRS and dexterity metrics. To assess inter-observer reliability of the GRSs the Cronbach alpha coefficient was applied to the 18 dual-observer rated videos. To assess intra-rater reliability the Cronbach alpha coefficient was applied to 12 re-scored videos. $p < 0.05$ was set as the limit of significance for all statistical tests.

Results

Global rating scales

Inter-rater reliability was determined by 2 blinded raters assessing 18 randomly selected videos from each group (table 1).

Intra-observer reliability was determined by the 2 blinded raters reassessing a further 12 randomly selected videos (**table 1**).

The Kruskal-Wallis test demonstrated significant differences between groups for the ASSET, BAKSSS and IGARS. Subgroup analysis showed that these differences were present between Novice vs. Trainee and Trainee vs Expert, demonstrating construct validity for all the GRSs (**figure 1, table 2**).

Correlations with dexterity metrics

Correlation of GRS scores with dexterity metrics were demonstrated to be both strong and statistically significant (**table 3**).

Discussion

With a number of GRSs now available for arthroscopic skills assessment, this first direct superiority assessment of such scales is timely. The three leading arthroscopic GRSs were used to assess arthroscopic performance during identical tasks on a VR simulator. Each GRS demonstrated excellent construct validity (ability to detect differences between skill levels), and each had excellent reliability, but none demonstrated superiority over another in differentiating between skill levels.

There is little prior work even within other specialities regarding direct comparisons of global rating scales. Aggarwal *et al* assessed three GRSs (one generic GRS, one a modification of this, and a task-specific GRS) alongside a task-specific checklist. This reported construct validity and strong inter-rater reliability for two GRSs, with the

procedure-specific assessments being inferior. This work did not however include a comparison between multiple different GRS.[23]

Each GRS demonstrated significant differences in ratings between skill levels of participants, and were maintained across all tasks irrespective of simulated joint (table **2**). For ASSET, BAKSSS and IGARS the greater difference was found between novice and expert participants (p<.001 for all tasks), than between trainee and expert participants (p<.001-.002 for all tasks). This is to be expected, as trainees have already made progression along a steep learning curve.

Previous work has demonstrated the validity of motion analysis in several surgical specialties, including orthopaedics and specifically arthroscopy.[10,24,25] The strong correlations shown by each GRS with the VirtaMed ArthroS™ VR dexterity output metrics recorded further strengthen the findings (table **3**).[26] Spearman's *rho* was found to be .757 - .952 for all tasks correlated with total time. Correlation with total path length were slightly lower in most tasks (Triangulation task of the knee remained *r*. 936 - .943), although the knee menisectomy tasks was notably lower at *r* of .641 - .673. This is likely due to multiple instruments being required, and as a consequence a higher overall path length as instruments are exchanged. **(CME2)**

The similarity of results between these GRSs is likely explained by the similarity of domains in each system. There are eight domains that are essentially the same between the GRSs, as presented in table **4**. Some of these domains are identical to the level of their Likert anchors.

Such similarities in both domain and anchors will unsurprisingly result in similar scores and thereby similar differences between skill levels. Whilst each GRS assessed has been developed independently, each includes core skill domains relevant to arthroscopy. The strengths and weaknesses of assessment for each domain are thus also the same, given the 5-point Likert scales and anchors. Whilst each GRS has attempted to improve these either with division of domains e.g. separate instrument and arthroscope handling domains, or with additional camera views as with the IGARS, it is arguable that these have made little difference. Prior work has also discussed the possibility that multiple domains assessing similar skills risk repetitive scoring.[27] This could potentially impact negatively on assessor attention, and thus the validity of their rating of a subject's performance.

Despite the potential for the perception of subjectivity in GRSs as assessment tools, they have been shown to be reliable. Each score was found to have a high degree of inter- and intra-rater reliability by their development groups, and these findings have been corroborated by the present study. Furthermore, an attempt to make a GRS 'more objective' with the inclusion of binary performance measures risks the score transforming into a task-specific check list style performance tool with their inherent limitations, as discussed earlier.

Anecdotally both assessors found each GRSs equivalent with regards to ease of use. This is unsurprising given the similarities between them (similar domains, same Likert-scale structure and similar written content). It was not felt that any one GRS could be recommended above the other with regards to usability.

This study has some limitations. The primary limitation was the modification of the GRSs with the exclusion of some domains. These were excluded on the basis that they were not relevant to the task in hand (e.g. instrument handling for arthroscope only tasks) or because they were not felt to be applicable to 'real-world' settings (e.g. exclusion of dual-camera system for IGARS recording both arthroscopic and body movements). Despite these pragmatic modifications in order to make each GRS suitable for everyday simulation skills use, they were still able to detect significant differences between each skill group indicating that these modifications did not affect the reliability or construct validity of the GRSs but also demonstrated the flexibility of these scales. This suggests the domains used represent the 'core' assessment domains for arthroscopy and would be applicable to the real world settings.

Whilst we have investigated the ASSET, BAKSSS and IGARS on a VR simulator, we feel our findings are applicable to GRS use in other forms of arthroscopic simulation. Bayona *et al.* developed the IGARS for use in 'both simulated and real settings', suggesting that the same skill domains are valid between different simulation formats. The ASSET and BAKSSS were initially designed for cadaveric simulation, but have since been used in other simulation formats and retain their validity. Finally, the highly similar performance across all three GRSs despite development for different simulation formats strongly suggests they can be utilised in bench-top, virtual reality and cadaveric arthroscopy simulation.

Finally, the GRSs assessed here are reported as being potential methods of assessment in the operating room. So far, these have only been critically assessed on simulations in skills centres. It is possible that the artificial nature of such an environment masks potential strengths or weaknesses of each GRS. Future studies should explore this concept in the operating room.

Conclusion

This study demonstrates that ASSET, BAKSSS and IGARS are equivalent tools for the assessment of performance on a virtual reality arthroscopy simulator. The modified GRSs were found to retain their construct validity, and these modifications have made them more applicable as assessment tools in simulated arthroscopy. However neither the ASSET, BAKSSS nor IGARS was superior over each other when used for the assessment of simulated arthroscopic skills. This is likely due to the inclusion of very similar domains in each of the rating scales. Furthermore, the similarities between these GRSs would suggest it is not worthwhile developing new GRS's unless they can address

skill assessment with more unique domains, and in a way that can be uniformly applied. (CME3) Consideration should be given to the adoption of a single GRS for the assessment of arthroscopic skills for ease of comparison in simulation research, and potentially in orthopaedic training.

References:

1. Landrigan CP, Rothschild JM, Cronin JW, Kaushal R, Burdick E, Katz JT, Lilly CM, Stone PH, Lockley SW, Bates DW, Czeisler CA. Effect of reducing interns' work hours on serious medical errors in intensive care units. *N. Engl. J. Med.* **351,** (2004).

2. Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J. Laparoscopic performance after one night on call in a surgical department: prospective study. *BMJ* **323,** 1222–1223 (2001).

3. Philibert I, Friedmann P, Williams WT. New requirements for resident duty hours. *JAMA* **288,** 1112–1114 (2002).

4. Benes V. The European Working Time Directive and the effects on training of surgical specialists (doctors in training): A position paper of the surgical diciplines of the countries of the EU. *Acta Neurochir. (Wien).* **148,** 1326–1330 (2006).

5. The implementation of the working time directive, and its impact on the NHS and Health Professionals. Report of the independent working time regulations taskforce to the Department of Health. (2014).

6. Temple J. Time for Training. A Review of the impact of the European Working Time Directive on the quality of training. (2010).

7. Dennis BM, Long EL, Zamperini KM, Nakayama DK. The effect of the 16-hour intern workday restriction on surgical residents' in-hospital activities. in *J. Surg. Educ.* **70,** 800–805 (Elsevier, 2013).

8. Pitts D, Rowley DI, Marx C, Sher L, Banks T, Murray A. A Competency Based Curriculum for Specialist Training in Trauma and Orthopaedics. *Br. Orthop. Assoc.* (2006).

9. Khera G, Milburn J, Hornby S, Malone P. Simulation in Surgical Training. A statement from the Association of Surgeons in Training. (2011).

10. Howells NR, Brinsden MD, Gill RS, Carr AJ, Rees JL. Motion Analysis: A Validated Method for Showing Skill Levels in Arthroscopy. *Arthrosc. J. Arthrosc. Relat. Surg.* **24,** 335–342 (2008).

11. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchinson C, Brown M. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br. J. Surg.* **84,** 273–278 (1997).

12. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: A systematic review. *Am. J. Surg.* **202,** 469–480 (2011).

13. Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, Stassen LPS, Dankelman J. Objective assessment of technical surgical skills. *Br. J. Surg.* **97,** 972–987 (2010).

14. Memon MA, Brigden D, Subramanya MS, Memon B. Assessing the surgeon's technical skills: analysis of the available tools. *Acad. Med.* **85,** 869–880 (2010).

15. Insel A, Carofino B, Leger R, Arciero R, Mazzocca AD. The development of an objective model to assess arthroscopic performance. *J. Bone Joint Surg. Am.* **91,** 2287–2295 (2009).

16. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad. Med.* **73,** 993–997 (1998).

17. Niitsu H, Hirabayashi N, Yoshimitsu M, Mimura T, Taomoto J, Sugiyama Y, Murakami S, Saeki S, Mukaida H, Takiyama W. Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surg. Today* **43,** 271–275 (2013).

18. Alvand A, Logishetty K, Middleton RM, Khan T, Jackson WFM, Price AJ, Rees JL. Validating a global rating scale to monitor individual resident learning curves during arthroscopic knee meniscal repair. *Arthrosc. - J. Arthrosc. Relat. Surg.* **29,** 906–912 (2013).

19. Koehler RJ, Amsdell S, Arendt EA, Bisson LJ, Bramen JP, Butler A, Cosgarea AJ, Harner CD, Garrett WE, Olson T, Warme WJ, Nicandri GT. The Arthroscopic Surgical Skill Evaluation Tool (ASSET). *Am. J. Sports Med.* **41,** 1229–37 (2013).

20. Olson T, Koehler R, Butler A, Amsdell S, Nicandri G. Is there a valid and reliable assessment of diagnostic knee arthroscopy skill? Knee. *Clin. Orthop. Relat. Res.* **471,** 1670–1676 (2013).

21. Bayona S, Akhtar K, Gupte C, Emery RJH, Dodds AL, Bello F. Assessing Performance in Shoulder Arthroscopy: The Imperial Global Arthroscopy Rating Scale (IGARS). *J. Bone Jt. Surgery. Am. Vol.* **112,** 1–7 (2014).

22. Fucentese SF, Rahm S, Wieser K, Spillman J, Harders M, Kock PP. Evaluation of a virtual-reality-based simulator using passive haptic feedback for knee arthroscopy. *Knee Surgery, Sport. Traumatol. Arthrosc.* 1–9 (2014). doi:10.1007/s00167-014-2888-6

23. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. *Ann. Surg.* **247,** 372–9 (2008).

24. Nasr A, Carrillo B, Gerstle JT, Azzie G. Motion analysis in the pediatric laparoscopic surgery (PLS) simulator: validation and potential use in teaching and assessing surgical skills. *J. Pediatr. Surg.* **49,** 791–4 (2014).

25. Mason JD, Ansell J, Warren N, Torkington J. Is motion analysis a valid tool for assessing laparoscopic skill? *Surg. Endosc.* **27,** 1468–77 (2013).

26. Datta V, Chang A, Mackay S, Darzi A. The relationship between motion analysis and surgical technical assessments. *Am. J. Surg.* **184,** 70–73 (2002).

27. Pellen M, Horgan L, Roger Barton J, Attwood S. Laparoscopic surgical skills assessment: can simulators replace experts? *World J. Surg.* **33,** 440–7 (2009).