Clinical Science

Coaching during a trauma surgery team training: perceptions versus structured observations

Alexander Alken, M.Sc.\(^{a,b,*}\), Edward Tan, M.D., Ph.D.\(^a\), Jan-Maarten Luursema, Ph.D.\(^a\), Cornelia Fluit, M.D., M.Sc.\(^b\), Harry van Goor, M.D., Ph.D.\(^a\)

\(^a\)Department of Surgery and \(^b\)Academic Educational Institute, Radboud University Medical Center, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands

KEYWORDS:
Coaching; Feedback; Faculty perception; Trainee perception; Non-technical skills; Team training

Abstract

BACKGROUND: Using the concept perception to quantify coaching skills during surgical training is questionable. This study compared the perceptions reported by the trainees and the faculty members following an emergency surgery team training with structured observations made on the basis of video registrations.

METHODS: For each faculty member, we scored 45 minutes of identical scenarios to enable the quantitative assessment of the use of positive feedback, corrective feedback, as well as instruction and in particular comments containing how and why explanations. We compared the values determined from the video scores with the perceptions reported by the trainees and faculty on questionnaires.

RESULTS: The trainee and faculty ratings for the coaching differed, with trainees generally giving a higher rating. While both the trainees and the faculty gave high ratings for the non-technical skills, corrective and complimenting feedback, and explanations why, the structured video observations showed lower scores in these categories.

CONCLUSIONS: Both the trainees and the faculty overrated the coaching. Trainee questionnaires and faculty self-reports neither reflected the actual coaching activities nor identified coaching skill deficits. © 2015 Elsevier Inc. All rights reserved.


The operating room (OR) is an important setting for learning surgical skills\(^1-4\) in which the faculty use real-time feedback and instruction to coach trainees.\(^5,7\) The teaching quality of the faculty is of interest to educational researchers and because research into surgical skills encompasses various aspects of teaching and learning, faculty self-reports and trainee questionnaires are considered necessary to examine the faculty’s teaching quality and to define areas for improvement.\(^4\)

Faculty self-reports and trainee questionnaires often differ from each other. Compared with trainees, the faculty usually report higher satisfaction with the amount, quality, and usefulness of real-time coaching\(^2\) and feedback.\(^2,8\) Regarding these differences in perceptions, it is not known whether the faculty members overestimate themselves or whether the trainees underestimate the faculty.\(^1,8\) When
using research that is based on perceptions to evaluate teaching, it is important to bear in mind that it is not known how close these perceptions reflect reality. Comparisons between perceptions and objective, structured measurements of teaching are required in order to increase the insight into the accuracy of the perceptions of faculty as well as trainees. Studies that have compared perceptions with structured measurements are primarily focused on how self-reporting residents assess their technical skills. Most studies conclude that self-evaluations are unreliable and that there are not any or only weak relations between self-evaluations and objective measurements like the Objective Structured Assessment of Technical Skills. However, some studies do show stronger relations and an improvement in accuracy with increased experience. For non-technical skills, both residents and more experienced surgeons proved not to be accurate when estimating their own performance.

Each year, under the supervision of the Dutch Trauma Society, the Radboud University Medical Center organizes the Definitive Surgical Trauma Care (DSTC) course: an acute surgical trauma team training on live porcine models during which trainees learn damage control techniques while supervised by experienced instructors. In the previous 2 editions, the exercises were video and audio taped to permit evaluation of the coaching activities. This provided us with the opportunity to investigate whether trainee and faculty perceptions of coaching agreed with each other. In addition, we could investigate the relationship between these perceptions and the observed findings of coaching on technical and non-technical skills that actually occurred in practice.

Methods

Participants

The surgical faculty consisted of 11 members. All members were experienced trauma surgeons with 2 to 12 years of teaching experience in trauma surgery. All of them had completed the Advanced Trauma Life Support train-the-trainer course, which includes learning the principles of effective feedback and giving instruction. Fourteen attending surgeons and 7 final-year registrars participated as trainees. All of them worked in hospitals that are involved in resident and/or medical student education. All trainees had limited or no experience in the treatment of the abdominal and thoracic trauma, the topic of this particular DSTC course. During this simulated practical emergency surgery team training, the faculty and the trainees operated in 11 teams. Each operating team consisted of 1 surgical faculty who coached 2 trainees, 1 scrub nurse and 1 member of the anesthesia faculty who coached 1 or 2 anesthesia trainees. Teams were composed in alphabetical order of last names. Only the surgical faculty and surgical trainees participated in this study. Written informed consent documents were obtained before the start of the training. As permitted by the Dutch law, the institutional review board of the Radboud University Medical Center waived the need for formal ethical approval.

Setting

Data were collected during the emergency surgery exercises, which are a part of the 3-day DSTC course. They took place at 4 ORs in the university’s central animal facility. The learning objective was to apply the principles and techniques of damage control surgery in scenarios using live, anesthetized porcine models with complex and multiple abdominal and thoracic injuries. Trainees were blinded concerning which injuries had been inflicted; both technical and non-technical skills had to be applied within a team setting. Within three and a half hours, the faculty guided the trainees through six acute scenarios. No formal instructions regarding effective coaching were provided. The local animal ethical review board approved this DSTC training.

Outcome measures

The faculty’s coaching behavior was recorded and analyzed using an observational coaching instrument (Table 1). Faculty self-reports and trainee questionnaires (Table 2) were taken to compare these 2 perceptions of coaching and to compare perceptions from both the faculty and the trainees with the actual coaching behavior that had been observed during the actual DSTC training practice.

Observational instrument

Coaching activities were recorded real time on audio and video to permit the observation and analysis of the coaching. A 45-minute segment containing 2 scenarios of the approximately 210 minutes recorded per team (6 scenarios) were used for this study: a stab wound to the infrarenal caval vein in the right mid abdomen and a pericardial stab wound to the left lung. These scenarios were selected because they occurred halfway through the training, thereby avoiding missing data by start-up difficulties in the 1st scenario and premature death of the animal in the last scenarios. The 2 scenarios did reflect the acute nature of the injury, the team setting, and the need for technical and non-technical skills, essential components of the complete training.

The faculty’s coaching behavior was categorized using a slightly modified version of a previously developed observational instrument: feedback was divided into either corrective or complimenting feedback (Table 1). For each faculty, a trained observer scored the amount of coaching activities and determined the type of reinforcement, ranging from least specific to extremely specific, accompanying the instructions or feedback concerning the team’s technical, communication, and team co-operation skills: not reinforced at all (least specific), reinforced with an
explanation how (moderately specific), an explanation why (highly specific), or both how and why combined (extremely specific). If one sentence addressed more observational categories, then each was scored and independently counted. Potential or ideal opportunities for coaching were not registered. The observer had been trained to observe and analyze recordings during the other 4 DSTC scenarios. In addition, the categorization was “optimalized” in a repeated manner during conferences with 2 experienced surgical educators and 2 experienced educationalists until face validity was considered good.

**Faculty self-reports and trainee questionnaires**

The faculty self-reports and the trainee questionnaires were designed to determine the perceptions of the coaching quality during the complete training. All questionnaire items were repeatedly discussed with the same surgical educators and educationalists until consensus was reached regarding the clarity of the formulated items and the face validity. The faculty self-reports and the trainee questionnaires consisted of 12 items: 9 of which were similar (Table 2, items 1 to 9) and the remaining 3 were comparable (Table 2, items 10 to 12). A 6-point scale was used to rate agreement with each questionnaire item (1 = absolutely not/never, 6 = absolutely yes/all the time) (Table 2).

Each faculty and trainee filled in the questionnaire independently from the other participants; this occurred immediately after the DSTC training had finished. Written instructions on how to fill in the self-reports and questionnaires were provided.

**Data analysis**

For each faculty member rating the 12 items on the self-report, 2 surgical trainees, who were coached by this faculty member, rated the same 12 items. To compare faculty with trainee perceptions, the total mean ratings per item of the trainees and that of the faculty members were calculated (Table 2).

It was possible to compare those items (measuring coaching quality) with the observational categories (quantity of coaching behaviors) by combining the observational categories for 8 of the 12 questionnaire items (Table 3).

The observed quality of the “corrective feedback” and the “complimenting feedback” was compared with items 2 and 5. By combining the observational categories “corrective feedback,” “complimenting feedback,” and “instructions,” the observed quantity of coaching on “technical,” “communication,” and “team co-operation skills” was compared with the questionnaire items 7 to 9. The categories “corrective feedback,” “complimenting feedback,” and “instructions” if “reinforced with only a how explanation,” “reinforced with only a why explanation,” or “reinforced with both a how and why explanation” were combined and compared with questionnaire item 10. The category “complimenting feedback” if “reinforced with a why explanation” or “reinforced with both a how and why explanation” were combined and the category “corrective feedback” when “reinforced

<table>
<thead>
<tr>
<th>COMPLIMENTING FEEDBACK</th>
<th>INSTRUCTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not reinforced at all</td>
<td></td>
</tr>
<tr>
<td>Reinforced with statement how</td>
<td>Reinforced with explanation why</td>
</tr>
<tr>
<td>Reinforced with statement how and explanation why</td>
<td></td>
</tr>
</tbody>
</table>

Technical skills

Communication skills

Team cooperation skills

CORRECTIVE FEEDBACK

Technical skills

Communication skills

Team cooperation skills
Variables and paired-sample Wilcoxon signed-rank tests for discrete and nonparametric whether variables were parametric or nonparametric. We used faculty group. Shapiro–Wilk tests were used to determine within-group effects were determined within the trainee and the between-groups effect. With the calculated mean scores between the mean scores of trainees and faculty to determine 11 individual faculty scores. Mean differences were calculated between the mean scores of trainees and faculty to determine why explanation’’ were combined and compared with items 11 and 12, respectively.

The mean scores for the faculty were calculated based on the 11 individual faculty scores. Mean differences were calculated between the mean scores of trainees and faculty to determine the between-groups effect. With the calculated mean scores within-group effects were determined within the trainee and the faculty group. Shapiro–Wilk tests were used to determine whether variables were parametric or nonparametric. We used Wilcoxon signed-rank tests for discrete and nonparametric variables and paired-sample $t$ tests for continuous and parametric variables. Correlations were calculated by Pearson’s $r$ (for continuous and parametric variables) and Spearman’s $\rho$ (for discrete and nonparametric variables). The comparison between the questionnaire items and the observed data was carried out by determining the correlations and by comparing the within-group effects from the questionnaires with the observed within-group effects. Effect sizes were calculated for $t$ tests (Cohen’s $d$) and the Wilcoxon signed-rank tests. An effect size of .10 or more is regarded to be small, .30 or more to be medium, and .50 or more to be a large effect. A $P$ value of .05 was considered significant. SPSS version 18.0 (SPSS, Inc, Chicago, IL) was used for all analyses.

### Results

Eleven self-reports from the faculty members and 20 trainee questionnaires were collected. Two trainees were excluded from the study because they had not returned the informed consent form. The internal consistency of the questionnaires was .87 (Cronbach’s $\alpha$ reliability). Table 2 shows the descriptive statistics of the mean ratings for the questionnaire and the observed frequencies for the available matching variables. Overall, the faculty (total questionnaire mean score: 4.5) and trainees (total questionnaire mean score: 5.0) rated satisfaction with the course.

### Trainee versus faculty perceptions

Trainee scores were higher for 11 of the 12 items compared with faculty scores (Table 2). The differences were significant for the items corrective feedback (item 2), timing of coaching (item 3), learning as a result of explanations why an action was incorrect (item 11), and learning as a result of explanations why an action was correct (item 12). The faculty rated themselves

### Table 2 Mean scores per item as perceived by the trainees and faculty (on a 6-point scale) and the available matching frequencies obtained when the coaching was observed during 45 minutes.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Trainees Mean</th>
<th>Trainees SD</th>
<th>Faculty Mean</th>
<th>Faculty SD</th>
<th>Observed frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trainees: I perceived the coaching by my faculty trainers as…</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. immediately applicable during the operation.</td>
<td>5.2</td>
<td>0.5</td>
<td>4.8</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>2. corrective, focused on what needed to be improved.</td>
<td>5.1*</td>
<td>0.7</td>
<td>4.6</td>
<td>0.7</td>
<td>2.1</td>
</tr>
<tr>
<td>3. well-timed.</td>
<td>5.3*</td>
<td>0.4</td>
<td>4.4</td>
<td>0.5</td>
<td>2.4</td>
</tr>
<tr>
<td>4. applicable in next situations.</td>
<td>5.2</td>
<td>0.5</td>
<td>4.7</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>5. complimenting, focused on what went well.</td>
<td>4.4</td>
<td>0.9</td>
<td>5.0*</td>
<td>0.4</td>
<td>2.3</td>
</tr>
<tr>
<td>6. exactly mounted to my personal needs.</td>
<td>4.6</td>
<td>0.7</td>
<td>4.2</td>
<td>0.8</td>
<td></td>
</tr>
</tbody>
</table>

**Faculty: I would describe my coaching of the trainees as…**

<table>
<thead>
<tr>
<th>Questions</th>
<th>Trainees Mean</th>
<th>Trainees SD</th>
<th>Faculty Mean</th>
<th>Faculty SD</th>
<th>Observed frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. technical skills.</td>
<td>4.8</td>
<td>0.6</td>
<td>4.6</td>
<td>1.2</td>
<td>12.3</td>
</tr>
<tr>
<td>8. way of communicating with the team.</td>
<td>4.7</td>
<td>1.1</td>
<td>4.4</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>9. way of cooperating with the team.</td>
<td>4.6</td>
<td>1.1</td>
<td>3.8</td>
<td>1.3</td>
<td>1.7</td>
</tr>
</tbody>
</table>

**Trainee: I learned a lot of the…**

<table>
<thead>
<tr>
<th>Questions</th>
<th>Trainees Mean</th>
<th>Trainees SD</th>
<th>Faculty Mean</th>
<th>Faculty SD</th>
<th>Observed frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>10. supplementary explanations and tips.</td>
<td>5.4</td>
<td>0.6</td>
<td>5.2</td>
<td>0.4</td>
<td>18.5</td>
</tr>
<tr>
<td>11. reinforcements why an action was correct.</td>
<td>5.2*</td>
<td>0.6</td>
<td>4.7</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>12. reinforcements why an action was incorrect</td>
<td>4.9*</td>
<td>0.7</td>
<td>3.9</td>
<td>0.9</td>
<td>1.7</td>
</tr>
</tbody>
</table>

*Trainee within-group effects: corrective feedback higher than complimenting feedback; mean difference = 0.7, $P = .02$, effect size = 0.5.

Effects within structured observations: more coaching on technical skills than communication skills; mean difference = 26.5, $t = 7.66, P = .000$, Cohen’s $d = 3.0$.

*more coaching on technical skills than team cooperation skills; mean difference = 27.9, $P = .003$, Wilcoxon signed rank test, effect size = 0.6.

*more coaching on communication skills than on team cooperation skills; mean difference = 1.4, $P = .02$, Wilcoxon signed rank test, effect size = 0.5.

*Trainees gave higher scores than faculty on: corrective feedback (mean difference = 0.6, $P = .04$, effect size = 0.4); timing of coaching (mean difference = 0.6, $P = .01$, effect size = 0.6); teaching following explanations why an action was correct (mean difference = 0.5, $P = .046$, effect size = 0.4); teaching following explanations why an action was incorrect (mean difference = 1.0, $P = .03$, effect size = 0.5).

The Faculty gave higher ratings than the trainees on: complimenting feedback (mean difference = 0.6, $P = .03$, effect size = 0.5).
significantly higher than trainees on giving complimenting feedback (item 5).

The faculty and trainee ratings on coaching for skills related to techniques, communication, and team cooperation were comparable between the 2 groups.

### Within-trainee-group effects versus within-faculty-group effects

Trainee ratings for corrective feedback were significantly higher than trainee ratings for complimenting feedback (Table 2, item 2 vs item 5), whereas the faculty ratings did not differ significantly. Trainees' ratings were not significantly different concerning coaching on technical, communication, and team cooperation skills. Similar results were found for the faculty ratings.

### Trainee and faculty perceptions versus structured observations

The trainees and the faculty gave high ratings for corrective and complimenting feedback and for explanations...
as to why an action was correct or incorrect. These ratings, however, did not agree with the observational findings (Table 2): the mean corrective and complimenting feedback was only used 2.1 and 2.3 times, respectively (item 2 and 5); explanations as to why an action was correct had not been used at all (item 11), while explanations why an action was incorrect had only been used a mean of 1.7 times (item 12).

Both the trainees and the faculty showed no significant differences for coaching on technical, communication, and team co-operation skills. This, however, did not agree with the scores of the video and audio observations (Table 2) in which the faculty significantly coached more often on technical skills (mean: 28.2 times; item 7) than on communication (mean: 1.7 times, item 8) or team co-operation skills (mean: .3 times, item 9).

A positive correlation ($\rho = .64$) was found for the faculty ratings and the observed number of coaching incidents for communication skills.

**Comments**

The faculty and the trainees both reported an overall high degree of satisfaction for the coaching during practical emergency surgery simulation training. Trainees, however, did rate the timing of that coaching, the corrective feedback, and being able to learn from specified instructions as being higher than the faculty. Both the trainee and the faculty perceptions of the coaching for technical, communication, and team co-operation skills differed from the real-time observation, demonstrating that the coaching on communication and team co-operation skills had been inflated in the perception data.

In contrast to most studies, trainees gave the faculty a higher rating than the faculty gave themselves. All trainees were attending surgeons or final-year registrars who worked in hospitals in which they taught junior residents or medical students. Possibly their own teaching experience made them understand the complexity of OR teaching and made them more lenient in rating the faculty’s teaching. Furthermore, senior trainees prefer training on live animal models above other simulations because of the realistic circumstances. This preference seems to be independent of the didactic course design or educational value.

Our real-time observational data indicated that the faculty overrated their coaching abilities on non-technical skills and the type and explanatory character of the coaching. The faculty may have rated their potential level instead of rating to the level at which they actually performed. It is also possible that the faculty lack appropriate insight into their coaching activities and thus overestimate their capabilities. However, Evans et al stated that the pressure to perform well and the need to impress is a more plausible explanation.

This study’s findings raise the question whether, in general, individuals are able to evaluate themselves with sufficient accuracy. The process of self-evaluation is complex and highly susceptible to contextual factors, therefore resulting in variable and inconsistent self-evaluations. Nevertheless, it is important to strive for accurate self-reports to establish the potential to learn and develop professional expertise. Video-based self-reflection has been shown to be an effective method to improve one’s ability to determine his own strengths and weaknesses. It may also be an effective method to improve one’s future teaching. In other research, however, external evaluations have been recommended to improve the reliability in identifying one’s strengths, weaknesses, and areas requiring improvement. The finding that those who have the lowest performance most often overrate themselves supports the need to use external evaluations. That no differences in overrating were apparent in this study is most likely because of the uniformly high rates with small standard deviations.

Evaluation by the trainees is considered necessary to examine the faculty’s teaching abilities and to identify those areas requiring improvement. We found that our trainees considered the faculty to apply important coaching activities that had rarely been observed during the video and audio analysis. For future evaluation, it seems necessary to 1st trainees in the criteria required to allow them to “objectively” evaluate courses and the faculty’s teaching abilities, especially when trainees and faculty may be involved in long-term training programs. Equipping trainees to be able to recognize effective and ineffective teaching most probably will improve future evaluations and will make it possible to identify areas that require improvement.

Although the need was stressed in 2002, more than a decade ago, valid tools to observe and analyze the quality of coaching during complex practical training are still not available. We have begun the process of validating our observational instrument in the DSTC and other emergency surgery courses. Our questionnaires were developed in close consultation with experts in medical and surgical education and have showed good internal consistency. The results obtained indicated that the perceptions were not in line with what we actually observed. While we selected 45 minutes from the operating scenarios, the trainees and faculty based their perceptions on the whole training. Therefore it remains possible that during other parts of the training the faculty did actually provide more coaching on non-technical skills. In our research, we compared quantitative observations with qualitative questionnaire outcomes of coaching perceptions. Even though a comparison has been made between quantitative and qualitative data, we think it is justified to conclude that these comparisons pointed out that the coaching and training would have been more effective if non-technical skills and detailed explanations would have been provided more often.
We compared the faculty’s self-perceptions with the trainee’s perceptions obtained by questionnaires to structured video observations of a practical surgical simulation team training for emergency trauma procedures to analyze the teaching as it occurred. Both the trainees and the faculty overrated the coaching activities, with trainees’ ratings being higher than faculty’s. This indicates that, at least in such emergency surgery courses, the use of trainee questionnaires and faculty self-reports will not provide adequate data to permit an analysis of the quantity and quality of the faculty’s coaching and to identify coaching skills deficits.

References