Assessment of Surgical Skills in Implant Dentistry

Vasileios Bousdras, DDS, MSc[1]/Behnam Aghabeigi, PhD, FDSRCS, FFDRCSI[2]/
Aviva Petrie, MSc, CStat[3]/Ann W. Evans, PhD, FDSRCS[4]

Purpose: *This study aimed to (1) compare 2 scales in the assessment of first-stage implant surgery, (2) assess the interrater reliability of these scales, and (3) compare self-assessment with observer assessment.* ***Materials and Methods:*** *Twenty-three patients underwent first-stage implant surgery. One assessor, an experienced dental surgeon, assisted and supervised the operator, while the second, a postgraduate trained in assessment, observed the procedure closely. The assessment scales consisted of a checklist and a global rating scale.* ***Results:*** *A significant correlation was found between the checklist and the global rating scale scores (r = 0.47,* P = *.002). The British Standards Reproducibility Coefficients were 2.5 (checklist) and 7.4 (global rating scale) for interrater reproducibility and 7.0 (checklist) and 12.6 (global rating scale) for self-assessment versus assessor reproducibility. Finally, analysis of the intraclass correlation coefficients between the assessors (0.74 and 0.64 for the checklist and the global rating scale, respectively) and between the surgeons' and trainers' scores (0.09 for the checklist and 0.18 for the global rating scale) showed a much weaker agreement for the latter.* ***Discussion:*** *There was good correlation between scores using the 2 different methods of assessment. The interrater reliability was substantial for both scales. However, training of assessors to ensure higher levels of interrater reliability may be necessary. These results also demonstrated the inability of some surgeons to assess their performance accurately.* ***Conclusion:*** *Both the checklist and the global rating scales provided useful assessment data, and both were considered of value by the assessors and surgeons in providing feedback. The development of assessment and self-assessment skills in implant surgery is necessary if we are to establish a culture of commitment to lifelong learning.* INT J ORAL MAXILLOFAC IMPLANTS 2004;19:542–548

Key words: *clinical competence, dental implants, educational measurements, outcome assessment, self-assessment*

[1]Honorary Clinical Research Fellow, Department of Oral and Maxillofacial Surgery, Eastman Dental Institute for Oral Health Care Sciences, University College London, United Kingdom.
[2]Consultant/Honorary Senior Lecturer, Department of Oral Surgery, Birmingham Dental Hospital, St Chad's Queensway, Birmingham, United Kingdom.
[3]Senior Lecturer, Biostatistics Unit, Eastman Dental Institute for Oral Health Care Sciences, University College London, United Kingdom.
[4]Senior Lecturer/Honorary Consultant, Department of Oral and Maxillofacial Surgery, Eastman Dental Institute for Oral Health Care Sciences, University College London, United Kingdom.

Correspondence to: Dr Vasileios Bousdras, Department of Oral and Maxillofacial Surgery, Eastman Dental Institute for Oral Health Care Sciences, University College London, 256 Grays Inn Road, London, WC1X 8LD, United Kingdom. Fax: +44 020 79151259. E-mail: V.Bousdras@eastman.ucl.ac.uk

This study was presented at the 16th Annual Meeting of the Hellenic Association of Oral and Maxillofacial Surgery (HAOMFS), November 30, 2002, Athens, Greece.

Dental implants provide a relatively easy means to achieve mechanical retention for a conventional removable denture or a fixed prosthesis and have gained worldwide popularity. The dental implantation procedure as first described by Brånemark emphasized the importance of an atraumatic, delicate surgical technique.[1] This method has been modified to simplify the procedure and reduce the treatment time.

Lambert and colleagues[2] proposed a minimum of 50 surgical implant placements as a guide for achieving basic competency in implant surgery. Over the past decade, as the number of clinicians who place dental implants has increased, the demand for formal training pathways and structured methods of assessment of surgical skills in implant surgery has gained recognition. Furthermore, clinical governance encourages assessment of competencies

Fig 1 Checklist scale for first-stage implant surgery.

	Incorrectly done/ not done	Done correctly
Surgical planning		
1. Preoperative assessment (radiographs, implant site, length)		
2. Patient preparation including anesthesia		
3. Appropriate design of flap		
4. Incision: length, depth, orientation		
5. Smooth reflection of flap in correct plane		
6. Soft tissue (and nerve) protection		
7. Evaluation of bony undercuts		
8. Preliminary location of implant sites (round bur and surgical guide if appropriate)		
Preparation of implant site:		
9. Angulation		
10. Depth		
11. Diameter (drill size)		
12. Countersink (if indicated)		
Implant placement		
13. Soft tissue retraction		
14. Implant position		
15. Torque		
16. Seating of cover screw/abutment		
Wound closure		
17. Single attempt at needle passage at correct height		
18. Follow-through on curve of needle		
19. Tying of knots		
20. Apposition of flap		

during training and after accreditation. Currently, logbooks and subjective assessment are the most popular methods by which technical competence in implant dentistry is evaluated. In spite of the agreement regarding the importance of competence in this area, all existing methods of evaluating surgical performance known to the authors are subjective.

The aim of this study was to determine whether a structured method of assessment could be devised. This study was carried out to *(1)* compare a checklist assessment scale with a global rating scale in first-stage implant surgery, *(2)* assess the interrater reliability of these scales, and *(3)* compare self-assessment with observer assessment.

MATERIALS AND METHODS

Participants

Staff, trainees, and postgraduates (8 surgeons in total) from various departments at the Eastman Dental Institute and Hospital in London were assessed while performing first-stage implant surgery under local anesthesia, with or without intravenous sedation. A total of 23 patients underwent surgery for the placement of 49 Brånemark System implants (Nobel Biocare, Göteborg, Sweden) by these 8 surgeons, following the protocol as originally described by Brånemark[3] with a few minor alterations. Most operations (19 of 23) were observed and assessed by 2 assessors out of a pool of 6 using both the checklist and the global rating scales. Assessor 1 was one of 5 staff members experienced in implant surgery; assessor 2 was a postgraduate (VB) who had been specifically trained in assessment techniques.

Study Design

Assessor 1 (staff member) assisted and, where necessary, trained the operator, while assessor 2 observed the procedure closely. Both assessment scales were shown to the surgeons prior to surgery and were completed immediately postoperatively by both assessors and surgeons. All surgeons were informed that this assessment was part of a research study and that it would not affect their final assessment.

Assessment Scales

Checklist Scale. The checklist scale (Fig 1) consisted of 20 important components of the first-stage implant surgery. The procedure was scored as correct or incorrect on each point; the total score possible ranged from 0 to 20. In cases where parts of the procedure were completed by the assessor-trainer, the relevant parts were judged as incorrectly performed.

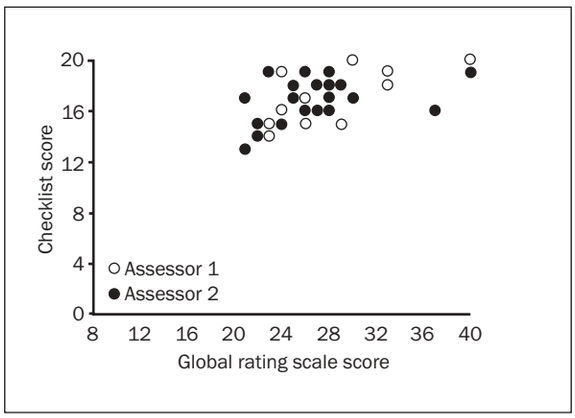

Fig 2 Scatter diagram showing the relationship between the checklist and the global rating scale (r = 0.47, P = .002) for both assessors.

Global Rating Scale. The global rating scale used in this study was that described by Martin and coworkers,[4] with minor modifications. This scale assessed surgical behavior and technique in general and has been shown to be a reliable tool for the assessment of surgical skills in general surgery.[5] It measured the following aspects: *(1)* respect for tissue, *(2)* time and motion, *(3)* instrument handling, *(4)* knowledge of instruments, *(5)* flow of operation, *(6)* use of assistants, *(7)* knowledge of the procedure, and *(8)* overall performance. Each of the 8 subscales was scored from 1 to 5, yielding a total possible score ranging from 8 to 40.

Statistical Analysis

The linear association between the 2 scales was measured by the Pearson correlation coefficient. The Bland and Altman method[6] was used to investigate the reproducibility of *(1)* the scores of the 2 assessors and *(2)* the score of the surgeon (ie, the self-assessment score) compared with the mean of assessors 1 and 2. Evidence of bias in investigations 1 and 2 was assessed by the paired t test. The result was considered significant if P was less than .05.

To measure the agreement between 2 assessors, the estimated standard deviation (SD) of differences provides a measure that can be used as a comparative tool. However, it is more usual to calculate the British Standards Reproducibility Coefficient (BSRC). This index measures the maximum likely difference between assessor 1 and assessor 2 and is calculated by the following formula[6]: BSRC = 1.96 SD. The BSRC was also used to measure the agreement between the score given by the surgeons (self-assessment) and the mean of the assessors' scores.

However, because the BSRC index for the checklist cannot be compared directly to the BSRC for the global rating scale, as they are based on different scoring systems, the intraclass correlation coefficient[7] was calculated. This index of reproducibility, often called the reliability index, lies between 0 and 1 and is closely related to the weighted kappa measure of agreement.

RESULTS

Correlation Between the Checklist and the Global Rating Scale

The Pearson correlation coefficient between the 2 scales was calculated and was shown to be statistically significant from zero (r = 0.47, P = .002) (Fig 2).

Interrater Reliability

As the mean difference in score for the checklist was estimated as 0.42 (95% confidence interval [CI] –0.21 to 1.05), and as this was not significantly different from 0 (t = 1.41, degrees of freedom [DF] =18, P = .18), there was no evidence of bias. There was also no evidence of bias for the global rating scale (mean difference 1.32, 95% CI –0.52 to 3.16, t = 1.5, DF = 18, P = .15). Moreover, using the Bland-Altman approach, there was no evidence of a funnel effect for either scale (Figs 3a and 3b), indicating that the difference in score between assessors did not vary with the magnitude of the score. The BSRCs for the checklist and the global rating scales were 2.5 and 7.4, respectively. Finally, the intraclass correlation coefficient was lower for the global scale (0.64) than for the checklist scale (0.74), which suggests more variability in scoring between the assessors for the global scale.

Level of Agreement Between Assessors and Surgeons

There was no evidence of bias for either the checklist (mean difference 0.52, 95% CI –1.2 to 2.25, t = 0.64, DF = 18, P = .53) or the global rating scale (mean difference 0.61, 95% CI –2.5 to 3.71, t = 0.41, DF = 18, P = .68). The Bland-Altman approach (Figs 4a and 4b) showed no evidence of a funnel effect for either scale; the BSRCs were 7.0 (checklist) and 12.6 (global rating scale). The intraclass correlation coefficients were calculated for self-assessment compared to the mean of assessors' scores for each rating scale. These coefficients were very low for both scales (0.09 and 0.18 for the checklist and the global rating scale, respectively) showing that scoring of some trainee surgeons compared to assessors was very different for both scales of assessment.

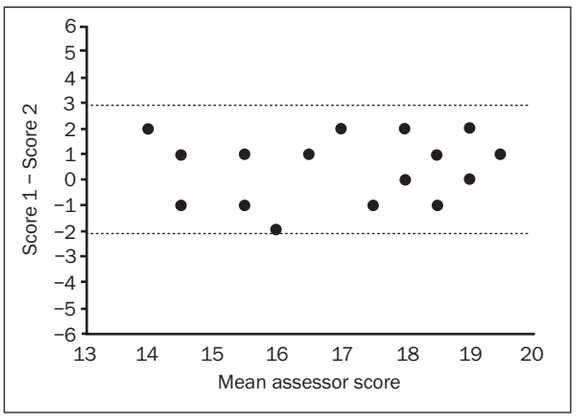

Fig 3a Differences in the checklist scores given by assessor 1 and assessor 2 plotted against their mean. Dotted lines indicate upper and lower limits of agreement. Score 1 = score given by assessor 1; score 2 = score given by assessor 2.

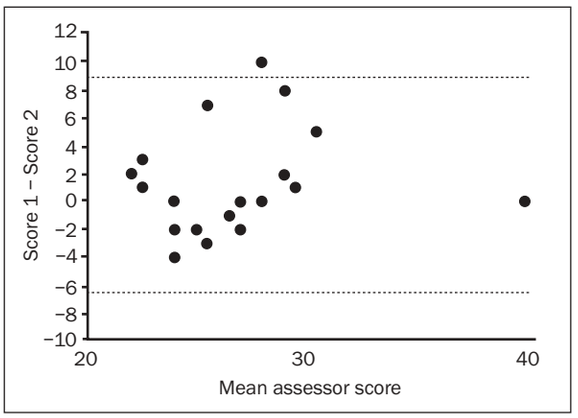

Fig 3b Differences in the global scale scores given by assessor 1 and assessor 2 plotted against their mean. Dotted lines indicate upper and lower limits of agreement. Score 1 = score given by assessor 1; score 2 = score given by assessor 2.

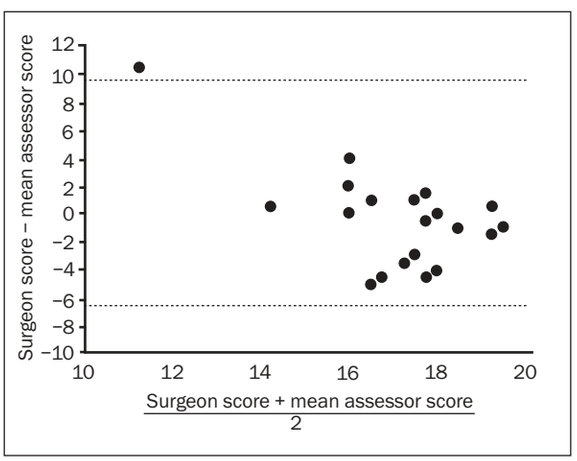

Fig 4a Differences in the checklist score given by the surgeon and the mean of the 2 assessors [(assessor 1 + assessor 2)/2] plotted against their mean. Dotted lines indicate upper and lower limits of agreement.

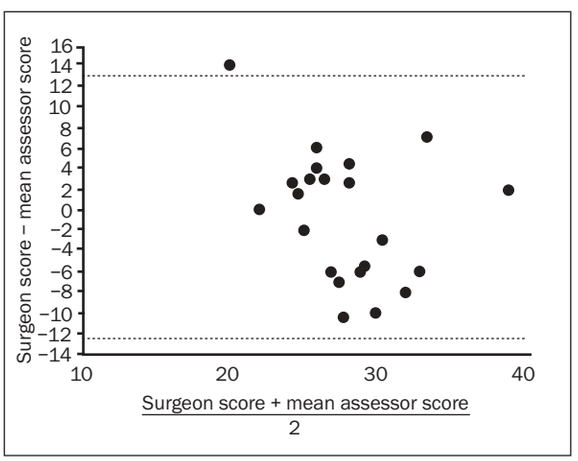

Fig 4b Differences in the global rating scale score given by the surgeon and the mean of the 2 assessors [(assessor 1 + assessor 2)/2] plotted against their mean. Dotted lines indicate upper and lower limits of agreement.

Moreover, a comparison of self-assessment scores against the scores given by both assessors in 19 cases (Table 1) showed that the percentage of times in which the 2 agreed was 27% with the checklist and 16% with the global rating scale.

Effect of the First Evaluation on the Second

Six surgeons performed at least 2 procedures; the other 2 surgeons performed 1 procedure each. To determine the effect, if any, of the first self-assessment on the second, the scores of the 2 evaluations were compared for each of the 6 surgeons who completed 2 or more self-assessments.

The checklist score increased for 2 surgeons, decreased for 3 surgeons, and did not change for 1 surgeon. The global rating scale score increased for 1 surgeon and decreased for 5 surgeons. The sample size was too small for a formal statistical analysis, but the available data suggested that the first self-assessment did not influence the second.

Similarly, to determine whether there were any changes in performance between the first and second operations by the same surgeon, a comparison was made of the mean scores given by 2 assessors for each operation for 4 surgeons who had performed at least 2 procedures (excluding 2 cases in which there was only 1 assessor).

The mean checklist score increased for 1 surgeon (ie, performance improved), decreased for 2 surgeons, and did not change for 1 surgeon. The mean global

Table 1 Agreement for Self-assessment Scores and Scores Given by Both Assessors in 19 of the Procedures

	Agreement		Overrating		Underrating	
	n	%	n	%	n	%
Checklist scale	5	27	9	47	5	26
Global rating scale	3	16	8	42	8	42

Table 2a Results of Evaluation Questionnaire A Completed by the 8 Surgeons

	Checklist		Global	
Questionnaire A	Yes	No	Yes	No
1. Did you feel the scales were of value in training/learning?	7	1	8	0
2. Did you feel these scales were of value in assessment?	6	2	7	1
3. Which scale did you think was the most fair?	3		5	
4. Which scale would be more valuable for giving feedback?	4		4	

Table 2b Results of Evaluation Questionnaire B Completed by the 8 Surgeons

Questionnaire B	Yes	No
1. Did you mind being assessed?	0	8
2. Did you feel that the assessment adversely affected your performance?	0	8
3. Did you feel any pressure to score yourself high or low and why?	1	7
4. Did you feel capable of self-assessing?	8	0

rating scale score increased for 2 surgeons (ie, performance improved) and decreased for 2 surgeons. Again, the sample size was too small for a formal statistical analysis.

Questionnaires

The 8 surgeons were given 2 questionnaires (Tables 2a and 2b) to evaluate their attitude toward assessment and their preferred method of assessment in relation to training and receiving feedback. Additional comments were invited.

All surgeons returned the questionnaires. All but 1 found both the forms of evaluation used to be of value in training and learning. However, their opinions were almost equally divided regarding their preferred scale. One surgeon felt pressured while completing the assessment forms, and another found it time consuming. All felt capable of self-assessment. Finally, half of them felt that a valuable format for feedback should include deficiencies and targets.

DISCUSSION

The authors set out to test whether the scales being used could measure competency in dental implant surgery in a more structured and reliable way. A good correlation between the results using the different scales would support (though not prove) the contention that the assessments measured the same quality[4] and could be applied to a range of levels of experience and training. Good agreement between the scores given by different assessors (ie, interrater reliability) would indicate whether the results could consistently be reproduced elsewhere and, hence, whether they were appropriate for "high-stakes" assessments. Finally, the level of agreement between the surgeon and the assessors might demonstrate the extent to which the scales would be helpful in self-assessment. Each of these aspects of the study is considered in the following discussion.

Correlation between 2 different scales would indicate whether they examine and measure the

same skills in a similar way. There was a statistically significant correlation between the checklist and the global rating scale in this study ($r = 0.47$, $P = .002$). However, higher levels of correlation between a checklist and a global rating scale have been reported in studies by Winckel and associates[5] and Evans and associates.[8] It would be surprising if there were no evidence of correlation in 2 scales purporting to measure the same ability, eg, surgical skill. However, Khan and coworkers[9] have suggested that some difference in results between the scales may be related to the fact that it is possible to score well in the task-specific assessment (ie, against fixed criteria) without being able to perform the task with the highest degree of skill. In addition, the global rating scale, but not the checklist, could take into account the surgeon's management of unexpected complications.[8] Because the checklist and the global rating scale measure slightly different skills, it was not surprising that identical degrees of correlation were not always seen between the scales. Checklist-type scales may be better for assessment at an early stage of skill development, eg, when a new technique is being learned, but the global rating scale is preferable for measuring proficiency as the skill is developed.

If assessment scales are to be used in high-stakes assessment, eg, qualifying examinations, it is important that there be good agreement between assessors or examiners. An intraclass coefficient (index of reliability) of at least 0.80 is usually considered desirable for such examinations.[4,10] In the present study the intraclass coefficient was substantial (0.74) for the checklist scale but lower (0.64) for the global rating scale. These results were supported by the BSRC. Agreement has been better in other studies, especially for the global rating scale,[5] which was previously found to be more reliable in this respect than the checklist approach.[11] The assessors in the present study were from a variety of dental disciplines and may have had different backgrounds, attitudes, and priorities in relation to surgical technique. This may partially explain the lower agreement between assessors. If these techniques were used in high-stakes assessment, it might be necessary to train assessors to ensure high levels of interrater reliability.

Poorer levels of agreement were found between surgeons and assessors than between assessors when assessing surgical skills in implant procedures. In addition, regarding both scales, surgeons' average scores were higher than their assessors' marks for them. The results of the present study are in line with those of a previous study, although the findings in that study were more pronounced.[12] Woolliscroft and colleagues[13] and Antonelli[14] have also suggested that weaker candidates tend to overrate themselves. Moreover, the results of the questionnaires suggested that some of the surgeons were unaware of how poorly they had assessed themselves. All felt that they were capable of self-assessment. The reasons for poor self-assessment could be summarized as follows: lack of understanding of what was expected,[12,15] scoring the potential or effort instead of actual performance,[13,16] an attempt to create a positive impression (fake good), and lastly, self-deception or lack of insight.[17] Surgeons were almost equally divided both on which scale was fairer and on which was better for providing feedback (Table 2a).

Clearly, none of the data invalidates the scales for use in assessment by others. However, the present data do demonstrate the inability of some surgeons to assess their performance accurately and reinforce the argument for teaching self-assessment skills during surgeons' formal training.

CONCLUSION

Implant surgery is a technically demanding and rapidly growing area of dentistry. Achieving a high degree of competence requires optimal training and assessment methods. Both the checklist and the global rating scales have been shown, in context, to provide reliable assessment data, and both were considered useful by the training surgeons in providing feedback. Furthermore, the scales could be used to compare different methods for teaching implant surgery. Significantly, some assessors went on to use the assessment scales in their practice for training of their junior staff and as a means of developing reflective practice. The value of the scales for self-assessment was limited by the lack of experience of the surgeons in self-assessment. The development of assessment and self-assessment skills in implant surgery is necessary if a culture of commitment to lifelong learning is to be implemented.

ACKNOWLEDGMENTS

The authors thank the patients, staff, trainees, and postgraduates who helped with this study. Ann W. Evans was supported by The Health Foundation. This work was undertaken by the authors in cooperation with University College London Hospitals, which receive a proportion of their funding from the National Health Service (NHS) Executive. The views expressed in this publication are those of the authors and are not necessarily those of the Trust or the NHS Executive.

REFERENCES

1. Albrektsson T, Brånemark P-I, Hansson HA, Lindstrom J. Osseointegrated titanium implants. Requirements for ensuring long lasting direct bone-to-implant anchorage in man. Acta Orthop Scand 1981;52:155–170.
2. Lambert PM, Morris HF, Ochi S. Positive effect of surgical experience with implants on second-stage implant survival. J Oral Maxillofac Surg 1997;55(12 suppl 5):12–18.
3. Adell R, Lekholm U, Brånemark P-I. Surgical procedures. In: Brånemark P-I, Zarb GA, Albrektsson T (eds). Tissue-Integrated Prostheses: Osseointegration in Clinical Dentistry. Chicago: Quintessence, 1985:211–240.
4. Martin JA, Regehr G, Reznick RK, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997;84:273–278.
5. Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. Am J Surg 1994;167:423–427.
6. Petrie A, Sabin C. Medical Statistics at a Glance. Oxford: Blackwell Science, 2000:93–95.
7. Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research, ed 4. Oxford: Blackwell Science, 2002:704–706.
8. Evans AW, Aghabeigi B, Leeson RMA, O'Sullivan C, Eliahoo J. Assessment of surgeon competency to remove mandibular third molar teeth. Int J Oral Maxillofac Surg 2002;31:434–438.
9. Khan KS, Bann SD, Campbell-Smith T, Darzi A, Butler PEM. Validation of surgical course assessment. Ann R Coll Surg Engl 2002;84(suppl):173–175.
10. Wass V, van der Vleuten CPM, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001;357:945–949.
11. Ault G, Reznick R, MacRae H, et al. Exporting a technical skills evaluation technology to other sites. Am J Surg 2001;182:254–256.
12. Evans AW, Aghabeigi B, Leeson R, O' Sullivan C, Eliahoo J. Are we really as good as we think we are? Ann R Coll Surg Engl 2002;84:54–56.
13. Woolliscroft JO, TenHaken J, Smith J, Calhoun JG. Medical students' clinical self-assessments: Comparisons with external measures of performance and the students' self-assessments of overall performance and effort. Acad Med 1993;68:285–294.
14. Antonelli MA. Accuracy of second year medical students' self-assessment of clinical skills. Acad Med 1997;72 (10 suppl 1):S63–S65.
15. Sullivan K, Hall C. Introducing students to self-assessment. Assess Eval Higher Educ 1997;22:289–303.
16. Arnold L, Willoughby TL, Cakins EV. Self evaluation in undergraduate medical education: The longitudinal perspective. J Med Educ 1985;60:21–28.
17. Evans AW, Leeson RM, Newton-John TR. The influence of self-deception and impression management on surgeons' self-assessment scores. Med Educ 2002;36:1095.